DOCUMENT RESUME

ED 442 844 TM 031 264

AUTHOR	Loomis, Susan Cooper
TITLE	Feedback in the NAEP Achievement Levels Setting Process.
SPONS AGENCY	National Assessment Governing Board, Washington, DC.
PUB DATE	2000-04-27
NOTE	56p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 25-27, 2000).
CONTRACT	ZA93003001; ZA97001001; RN91226001
PUB TYPE	Numerical/Quantitative Data (110) -- Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	*Academic Achievement; Academic Standards; *Civics; Classification; Cutting Scores; Elementary Secondary Education; *Feedback; *National Competency Tests; *Teachers; *Validity
IDENTIFIERS	*National Assessment of Educational Progress; *Standard Setting

ABSTRACT

This paper describes the feedback included in operational achievement level setting (ALS) procedures for the National Assessment of Educational Progress (NAEP). It does not describe the feedback form field trials, pilot studies, or other research studies related to the NAEP. The NAEP ALS process includes three rounds of item-by-item ratings, and feedback is provided after each round. This means that panelists have additional information to consider for subsequent rounds of ratings, but additional training is required for no more than one of two new forms of feedback for each round. Most of the feedback is aimed at increasing interjudge and intrajudge consistency. The types of feedback provided are: (1) cutscores and standard deviations; (2) student performance data; (3) interrater consistency data; (4) whole booklet data; (5) interrater consistency data, including "Reckase" charts of performance data; and (6) consequences data. The use of each type of feedback is discussed, and panelists' evaluations of the feedback provided are noted. Panelists' responses indicate high levels of understanding about and confidence in using the feedback. The Reckase charts appear to be a significant addition to the array of feedback information. (Contains 9 tables, 17 figures, and 22 references.) (SLD)

Feedback in the NAEP Achievement Levels Setting Process

by
Susan Cooper Loomis
NAEP ALS Project Director
ACT, Inc.

This paper was prepared for presentation at the National Council on Measurement in Education meeting, April 27, 2000, Le Meridien Hotel, New Orleans.

The research for this paper was supported in part by contract ZA97001001 with the National Assessment Governing Board. Research reported in this paper was supported by earlier contracts with the National Assessment Governing Board ZA93003001 and RN91226001.

Feedback in the NAEP ALS Process

Susan Cooper Loomis, ACT[1]

Introduction

ACT has been the achievement levels-setting contractor to the National Assessment Governing Board (NAGB) for nine years. This period can be divided into three larger groupings or cycles of setting achievement levels for NAEP. In the first cycle, achievement levels-setting procedures were implemented for mathematics, reading, and writing in 1992. The second included geography and U.S. history in 1994 and science in 1996. The third included civics and writing in 1998.[2] ACT has implemented achievement levels-setting (ALS) procedures for seven different subjects included in the assessment program of the National Assessment of Educational Progress (NAEP).

Judgments are a central feature of standard setting. Judgments must be well informed. One of the most important parts of the NAEP ALS process is feedback to panelists to inform their judgments. This paper will provide a review of various forms of feedback that have been provided by ACT during this period of setting NAEP achievement levels.

Webster defines *feedback* as "the return to the input of a part of the output of a machine, system, or process." In a production process, feedback is provided to increase the efficiency and quality of output. NAEP ALS panelists want to do a good job—to produce a good product, and they are always asking whether they did it *right*. Being told that judgments are neither right nor wrong leads to some frustration on the part of panelists. The goal has been to provide as much information to panelists as is deemed to be feasible. Feasibility is judged to be a function of several factors:

- the data that are available at the time of the ALS study,
- the data that are consistent with NAGB's policies on achievement levels,
- the data that can be produced on site within time constraints imposed by the agenda,
- the data that can be understood by panelists after an amount of training consistent with time constraints imposed by the agenda, and
- the amount of data that panelists are able to consider.

ACT is increasingly adding feedback to the training portion of the ALS process to help inform panelists about their performance judgments during training exercises. The feedback to be described in this paper, however, is that related only to the rating process, per se, and not to the exercises specifically related to training. Further, this paper describes only the feedback included in operational ALS procedures. Feedback used in field trials, pilot studies, and other research studies related to the NAEP ALS process are not described here.

---

[1] I wish to acknowledge the assistance of Teri Fisher in preparing this paper. She developed the tables and figures from materials previously reported in various documents.

[2] After analyzing the evidence collected from the 1992 Writing NAEP Achievement Levels-Setting Process, NAGB decided against setting achievement levels for writing until the test specifications were redesigned.

Overview

The NAEP ALS process includes three rounds of item-by-item ratings, and feedback is provided after each round[3]. The general strategy has been to provide additional data after each additional round of ratings. Most of the feedback provided after Round 1 is updated and provided again after each subsequent round. Some additional types of feedback are added after subsequent rounds, and some are deleted in subsequent rounds. This means that panelists have additional information to consider for subsequent rounds of ratings, but additional training is required for no more than one or two new forms of feedback for each round.

NAGB's policy on NAEP achievement levels specifies that the method for collecting panelists' judgments must be criterion-referenced. In addition, Congress has mandated in Public Law 103-382 that NAGB show that the achievement levels are "useful, reasonable, and valid." It is likely that panelists engaged in a five-day process will want some information to judge for themselves whether the outcomes of their efforts seem useful, reasonable, and valid. This means that some normative information must be provided to panelists although the major impact of the feedback should not be normative.

ACT's goal has been to meet the diverse, and somewhat conflicting, demands set for NAEP achievement levels. ACT has been guided in this endeavor by some of the very best experts on standard setting. The NAEP ALS Project Staff are advised by the Technical Advisory Team (TAT)[4] and the Technical Advisory Committee on Standard Setting (TACSS). The TACSS meets approximately every two months, although meetings are scheduled around project activities so that some are scheduled more or less frequently than that. Past and current members of TACSS are as follows.

William Brown, Brownstar, (1993 – present)
Barbara Dodd, University of Texas-Austin (1997 – present)
Robert Forsyth, University of Iowa (1991 – present)
Ronald Hambleton, University of Massachusetts-Amherst, (1991 – present)
Michael Kane, University of Wisconsin (1991 – 1997)
Brenda Loyd (deceased), University of Virginia (1991 – 1995)
John Mazzeo, ETS[5] (1997 – present)
William Mehrens, Michigan State University (1991 – present)
Jeff Nellhaus, Massachusetts department of Education (1997 – present)
Mark Reckase, Michigan State University (1998 – present)[6]
Douglas Rindone, Connecticut Department of Education (1993 – present)
Wim van der Linden, University of Twente (1997 – present)
Rebecca Zwick, University of California-Santa Barbara (1997 – present)

TACSS provides recommendations regarding all aspects of the ALS process. They evaluate all study designs and make recommendations to improve them. They evaluate all analyses of ALS procedures and make recommendations regarding their interpretation as well as recommendations for modifications of

[3] A "final round" was added to the 1998 NAEP ALS process. Panelists selected a cutscore to represent performance at the lower boundary for each achievement level. The values selected were to represent the individual panelist's cutscores, and they were averaged to form the grade-level cutscore. The final cutscores and the consequences associated with each, i.e. the percentages of students performing at or above each, were reported as feedback to this final round.

[4] The present membership of the Technical Advisory Team includes Nancy Petersen, Richard Sawyer, Bradley Hanson, Catherine Welch, and Robert Brennan.

[5] One representative of ETS serves on the TACSS. From 1991 – 1995, Eugene Johnson was the representative. From 1995 – 1997, James Carlson was the representative.

[6] Mark Reckase served on the TAT from 1991 until he joined the faculty of Michigan State University in 1998.

and additions to the analyses. The feedback described in this report have been developed and modified through input from the TAT and TACSS.

Feedback in the NAEP ALS Process

The major aim of feedback provided in the NAEP ALS process has been to inform the judgments of panelists to maximize the quality of the output of the process. "Consistency" is generally held to be an important indicator of the quality of a standard setting process. Most feedback has been aimed at increasing interjudge and intrajudge consistency. Beginning with ACT's first proposal in 1991 to set NAEP achievement levels, interjudge and intrajudge consistency have been considered as two of the most important criteria by which to judge the ALS process. Although the format and specific types of feedback have changed over time, most of the feedback provided in the NAEP ALS process has primarily focused on these two features of a standard setting process. The NAEP feedback has not been limited to these two categories of information, however. In addition, ACT provides student performance data for each item and for whole test booklets. The NAEP ALS process includes a variety of information for panelists.

Beginning with the 1998 ALS process, ACT decided on a specific ordering of the feedback data presented to panelists. Prior to that, the order was less intentional. For each type of feedback, an example is provided, along with an explanation about the source of the data, instructions about how to interpret the data, and instructions about how to change item ratings in response to the information. The feedback data are first presented in a general session to all panelists in all grades, and instructions in their use are provided in the general sessions. The feedback data are generally distributed in grade-level groups where panelists discuss them with one another and receive individual assistance from the grade-level process facilitators. These types of feedback are currently provided in the NAEP ALS process:

1. Cutscores and standard deviations
2. Student performance data
3. Interrater consistency data (rater location charts)
4. Whole booklet data
5. Reckase Charts (intrarater consistency data)
6. Consequences Data

This sequence will be followed in the descriptions of feedback provided below. Next, information about panelists' evaluations of the feedback will be presented for each of the procedures in which it has been provided. The same feedback data were provided for each of the subjects for which achievement levels were set within an administration cycle. Panelists' evaluations of feedback will be compared across the subjects within the same ALS process cycle, and they will be compared across subjects and cycles when possible. First, it is useful to have a clear understanding of each type of feedback.

Cutscores and Standard Deviations

ACT presents the cutscores and standard deviations for each achievement level at each of the three grades. The meaning of a standard deviation is briefly described in very simple terms. Logical patterns to be expected are described and explained to panelists. They are told whether the data are scaled on a within-grade or an across-grade scale and the implications of that fact for their interpretations. Figure 1 is an example of the 1992 version of this feedback data reporting cutscores and standard deviations, and Figure 2 is an example of the version used since 1994. An across grade scale was used for the subjects included in the 1992 ALS process. The 1998 ALS cycle represents the first for which student performances were reported on a "purely" within-grade scale.

Student Performance Data

ACT provides data to help provide panelists with a "reality check." After the first round of ratings, ACT provides a complete set of student performance data for each item. These data are not feedback, in that the process does not produce them, but they are typically provided in a modified Angoff rating process after the first round of item ratings. For multiple choice and other dichotomously scored items, the data are p-values, i.e., percentages of students with correct responses. For constructed response items, i.e., polytomously-scored items, the mean score is reported along with the percentages of students scoring at each rubric score value (e.g., 1 = not correct; 2 = partially correct; 3 = completely correct), not responding, or not receiving credit for their response.

In the 1992 Mathematics NAEP ALS process, student performance data were modified so that the percentages of student responses scored 1-4 summed to 100%. This eliminated the percentages of students who gave no response or who gave a response judged to be "off task." The decision to recompute the percentages was made without knowledge of the fact that many, many students in the 1992 Math NAEP either did not respond to the constructed response items or gave responses that were "off task." That method of recomputing the percentages was not again implemented. An example of student performance data tables provided to panelists in the 1992 Math ALS process is presented in Figure 3, and an example of the student performance data provided in the 1992 Writing NAEP ALS is presented in Figure 4. Note that all items in the Writing NAEP are polytomously scored items. Figure 5 shows an example of the format used since then.

Interrater Consistency Data

Interrater consistency feedback is presented in charts with histograms. These charts are referred to as "rater location charts." Interrater consistency or rater location data show where each rater's cutscore was set for each of the three achievement levels. Panelists can compare their own cutscore to that of every other panelist in the grade group and to the overall cutscore for the grade group.

Letter codes identify the location of each rater at each level. These letter codes are given to panelists as "secret codes." Panelists tend to share their secret code identity with one another immediately, but that is their choice. Panelists who wish to remain anonymous, with respect to their rater location, may do so.

Rater location charts show the cutscore locations of specific panelists for specific achievement levels. An example of the charts from 1992 is presented in Figures 6.1 – 6.3, and an example from the 1998 Civics ALS process is presented in Figures 7.1 – 7.3. On the rater location chart for the Proficient level, for example, the letter codes reveal the specific location of the Proficient level cutscore for each panelist in a grade level. On that chart, panelists see shading where the Basic and Advanced cutscores were set. This feature allows panelists to determine whether any Proficient cutscore, for example, was set in the range of Basic cutscores or Advanced cutscores. The letter code allows the panelist to determine where his or her own cutscores were located relative to the overall grade level cutscore and relative to all other panelists in the grade group.

The mean and standard deviations of the ACT NAEP-like scale have changed since the 1992 process, but the basic format of the interrater consistency data has not changed. There have been instances, such as the 1998 field trial for writing, when the range of cutscore values on the ACT NAEP-like scale was too great to be represented on an 11" × 8.5" page. Cutscores were marked at either the lowest or highest value, and panelists were notified that their actual location could be much lower or higher than the value on the chart.

Whole Booklet Data

There are actually two forms of whole booklet feedback data. One is referred to as an "exercise" and the other as feedback data. Panelists participate in the exercise only one time, as part of the review of feedback from Round 1 ratings. The whole booklet feedback is provided after each round of ratings.

This whole booklet feedback was added to the process in 1994 in response to suggestions made by the National Academy of Education (NAE) Panel in their evaluation of the 1992 process. (NAE, 1993). This information is a holistic version of the student performance data described above. Panelists take a form of the NAEP on the first day of the ALS process. Two forms are administered at each grade. Panelists do not rate the items included in the NAEP form administered to them, but they do get feedback about student performance on those booklet forms.

For the exercise, panelists are given up to four booklets that are scored within 2% (i.e., ±2%) of the total percent of possible points needed for performance at the cutscore of each achievement level. The cutscore used for this exercise is the grade level cutscore set by Round 1 ratings. The booklets are selected from a set of 100 booklets randomly selected from that form. Since the booklets are randomly selected, there is no guarantee that there will be any booklets that meet this criterion. If not, then no booklets are presented to panelists to illustrate the performance of students at a particular level. If there are several booklets from which to choose, ACT staff select booklets to represent as much diversity as possible with respect to alternative response patterns resulting in the requisite total score. There are no scores marked on the booklets or on the individual items. Booklets are identified according to the achievement level matching the score (percentage of total possible points). If the booklet score is exactly the percentage of total possible points associated with the cutscore, then the booklet is marked as B (Basic), P (Proficient), or A (Advanced). If the booklet score is slightly higher than the total points at the cutscore (maximum of +2%), the booklet is marked with the level and a "+." If the booklet score is slightly lower than the total points at the cutscore (maximum of –2%), the booklet is marked with the level and a "–." Since panelists took that particular NAEP form and scored their own responses, they were already familiar with the items and scoring rubrics for the items. Panelists are given the scoring keys and rubrics for items in the form, and they can refer to those during their evaluation of the performance of students in each booklet they review.

Panelists are given complete instructions regarding the sample of booklets available, how the scores are marked, and how to interpret the information. Instructions to panelists stress the "holistic" nature of the exercise and urge panelists to read all of the responses to form an overall evaluation of the level of performance represented by the responses. If the performance seems "higher" than expected for borderline Basic performance, for example, then panelists should consider lowering ratings for borderline Basic. If the performance seems "lower" than expected for a level, then panelists should consider raising ratings for that level.

Panelists seem to understand these data, although some panelists are extremely frustrated by the lack of scores for each item in the booklets.

The whole booklet feedback is based on the same NAEP booklet form. Panelists are simply informed about the percentage of total possible points needed to qualify for performance at the cutscore of each achievement level. The data are updated after each round of ratings, and the data are provided to panelists for their consideration regarding changes to make in ratings in the following round. These data provide information regarding both the direction and magnitude of change desired.

The format of these data has changed somewhat since the first uses in 1994. Examples of the various formats and descriptive statements presented to panelists are included in Figures 8-10.

Panelists seem to understand the data and to take it into account in making their decisions regarding whether to change their ratings to increase or decrease their cutscore for a particular achievement level. During the 1996 Science ALS process, one grade 8 panelist tried to convince the other panelists in the grade group to lower cutscores as a result of these data. She appealed to the other panelists to think honestly about their own performance on the assessment and to judge whether they had performed as well as the requirements they had just set for students to meet the Advanced level. She pointed out that as adults, most of the panelists had much more training and experience in science than the students. Although she was not successful in convincing the panelists to lower their cutscores, she succeeded in convincing the facilitators that she really understood what the data meant and how to use the feedback for the next round of ratings.

Intrarater Consistency Data

Informing panelists about the consistency of their own ratings has been the most troublesome task of providing feedback to NAEP ALS panelists. ACT experimented with several different formats before deciding to eliminate this feedback altogether from the 1996 Science ALS process. The problem was to determine some reasonable criteria to guide the provision of intrarater consistency data to panelists. A minimum level of inconsistency could be set that would assure that all panelists would have ratings to re-evaluate for at least one or two items on the feedback list. On the other hand, using a really low level as the criterion would potentially result in some panelists having all of the items listed for re-evaluation. Similarly, listing the same number of items for each panelist would potentially lead to the interpretation by panelists that all of the ratings for items on the list should be changed and that only the ratings for items on the list should be changed. Panelists with relatively high levels of intrarater consistency would potentially perceive that their consistency was no higher than panelists with relatively low levels of intrarater consistency, and vice-versa. Figures 11 and 12 provide examples of these efforts at providing intrarater consistency feedback to panelists during the 1992 and 1994 NAEP ALS cycles.

Reckase Charts were proposed as part of a new rating method. The rating method was tested in field trials, and the final decision was to use the charts along with the same rating method used since 1994 (Loomis, et al., 1999). Reckase Charts seem to have solved the dilemma regarding how to provide intrarater consistency feedback that can be easily understood and used by panelists. Reckase Charts contain data on student performance based on estimates from the Item Response Theory model used by NAEP. Data for each item are presented in columns on a chart. Each column presents data for one item. Rows represent performance estimates across each item for students performing at each score point. A row contains expected student performance data for students scoring at that point on the NAEP scale. An abbreviated version of a Reckase Chart is presented as Figure 13.

Panelists can evaluate the consistency of their ratings with respect to several item attributes. In particular, panelists can evaluate their ratings for multiple choice and other dichotomously scored items relative to their ratings for constructed response/polytomously scored items. They can evaluate their ratings for items in different content domains of the overall subject to determine whether they have been consistent in their ratings of life and earth science items, for example, or persuasive and informative writing tasks. Panelists can even evaluate the consistency of their ratings across the different item blocks to determine whether they became more consistent as they gained experience by rating more items or, perhaps, less consistent as they became more fatigued.

Consequences Data

Feedback referred to as "consequences data" are provided to inform panelists about the results of the ALS process. In this case, the "results" are not the "outcomes" of the ALS process, such as the cutscores, items, and descriptions of the achievement levels. Rather, these consequences result from the cutscores set by ALS panelists in combination with the performance of students on NAEP. The consequences data provided to NAEP ALS panelists are the percentage of student scores at or above each cutscore. Panelists

are given this information to help them evaluate whether the cutscores they have just set through the item rating process seem reasonable.

No consequences data were presented to panelists during the 1992 ALS process. Beginning with the 1994 ALS process, panelists were given consequences data ***after*** the final round of item ratings. Panelists were instructed in the concept of "at or above," and consequences feedback was reported as the percentage at or above each level. See Figure 14 for an example of grade-level consequences data reported in 1994 and Figure 15 for an example of the current format which was introduced in the 1996 Science ALS Process. Starting with the 1996 Science ALS, a pie chart was included to report the percentage of students performing *within* each achievement level category, plus the percentage of students performing below the Basic level. The electronic version of the graphic presented to all panelists in a general session included a modification to represent more clearly the percentage *at or above* each level.

The information was shared with panelists as feedback, and their reactions to the data were collected to share with NAGB. Grade-level consequences data were shared with all panelists in a general session, so all panelists in each grade were aware of grade-level differences. A copy of the questionnaire developed to collect panelists' reactions to the consequences data is included as Appendix A to this report. The questionnaire has been modified only slightly to fit the circumstances of administration. Panelists are asked whether the consequences data meet their expectations, whether they would want to make some changes, and which cutscores they would want to change. For each achievement level cutscore, panelists are given three options.

- Raise the cutscore to lower the percentage of students scoring at or above the level
- Lower the cutscore to raise the percentage of students scoring at or above the level
- Make no change

The responses of panelists generally indicated that few would change cutscores, and the changes that were recommended were generally rather small.

The Governing Board was reluctant to approve the provision of consequences data to panelists during the rating process because they wanted to have the NAEP achievement levels cutscores based on the achievement levels descriptions and not on the performance of students. In discussions of this issue in 1996, the Achievement Levels Committee concluded that consequences data were to be considered by NAGB when making the decision on where to set the cutscores—not data to be consider by panelists when developing cutscores to recommend to NAGB. Nonetheless, NAGB seemed to pay special attention to the responses of panelists regarding the consequences data. Those responses seemed to have had a particularly important impact on the decision of some Board members to accept the U.S. history cutscores in 1994, for example.

NAGB approved ACT's proposal to study the impact of consequences data on cutscores as part of the current contract awarded in 1997. The design of field trials for the 1998 ALS process included the provision of consequences data to one group after each round of ratings and to the other only after the final round of item ratings. Both groups of panelists were all allowed to recommend different final cutscores after seeing the consequences data for the final round of item ratings.

The findings from the field trials were that consequences data had little impact on the cutscores set by panelists. Panelists who received consequences data throughout the process were not more likely to set higher or lower cutscores than panelists who received consequences data only at the end of the rating process.

NAGB agreed to have ACT provide consequences data to panelists after the final round of item ratings. The panelists would be asked to recommend a final cutscore for each achievement level for their grade, after having reviewed the consequences data. These final recommendations were to be used to compute the final cutscore to be recommended to NAGB. This plan meant that NAGB could select the cutscores set after the three rounds of item ratings—before panelists were given consequences data—or the cutscores recommended by panelists after they were informed about the consequences of their ratings.

The results of the 1998 pilot study revealed a pattern of change that had not been evidenced in any previous NAEP ALS process. The cutscores increased for each level at each grade and across every round. After reviewing the results, TACSS recommended that consequences data be provided after Round 2. The Reckase Charts were new additions to the process, but there was no way to know whether this finding was somehow associated with the use of the Reckase Charts in the process. There was no way to know whether this pattern would ever be repeated. If panelists raised their ratings after being informed about the consequences of cutscores based on their ratings, then there was some assurance that this pattern was intended. At the very least, the consequences data would allow panelists to make informed decisions regarding the direction of change they wanted to implement in their ratings.

NAGB agreed to allow consequences data to be provided to panelists during the process. The recommendations of TACSS were followed for the Writing NAEP Pilot Study and for the ALS process in both civics and writing. Panelists were given grade level consequences data following Round 2 ratings. Figure 15 is an illustration of the grade-level data shared with panelists. These data were presented to panelists in a general session so all panelists were informed about the consequences data for cutscores at all levels and all grades. Panelists were given instructions about how to interpret the data and how to change their ratings to increase or decrease the percentages of students at or above a cutscore. Panelists had an opportunity to discuss the consequences data in grade groups as part of their overall review of Round 2 feedback.

Following Round 3 ratings, panelists were again given consequences data for their grade level cutscores. These data were again provided in a general session to panelists in all three grades. In addition to the grade-level consequences data, panelists were given feedback about the consequences of their own cutscores. Consequences graphs showing the percentages of students scoring at or above each level for the panelist's grade were again shared with panelists. The Round 2 feedback forms were updated to reflect Round 3 cutscores. The rater location charts used to provide panelists with interrater consistency feedback were modified to provide panelists with more consequences data. The rater location charts were marked to show the percentages of students scoring at or above various score points. The consequences data were reported on the rater location charts at 5-point increments on the scale. Please refer to Figures 16 and 17 for examples of these feedback forms.

The real difference in consequences data provided after Round 3, however, was that individual-level consequences data were given to panelists. For each grade group, a list with each panelist's "secret ID" was distributed to each panelist. The list provided the cutscores and consequences data associated with each for the three achievement levels. Panelists were asked to evaluate the information and to complete a consequences data questionnaire. On the questionnaire, they were to recommend a final cutscore for each achievement level. They were instructed that their recommended cutscores would be used to compute an overall grade-level cutscore. The overall grade-level cutscore would be the final cutscore used for selecting exemplar items to use in reporting, and ACT would recommend that set of cutscores to NAGB—unless some evidence indicated another course of action.

Panelists recommended changes to 40 cutscores for civics and 53 cutscores for writing (Loomis, 1999b; 1999c). Those numbers represent about 15% of the total number of cutscores in civics and 20% in

writing. If a panelist recommended no changes to cutscores, the cutscores from Round 3 were used in computing the average for the final cutscores. Panelists were informed that this would be the procedure.

After the new cutscores were computed, based on individual recommendations, consequences data were updated to reflect these final cutscores. Grade-level consequences data were again shared with all panelists in a general session. A consequences data questionnaire was again administered. Panelists were asked the same questions. They were told that the cutscores would not be recomputed on the basis of recommendations for change. ACT explained that their responses and recommendations would be shared with the technical advisors and with NAGB. If the evidence collected from the questionnaires indicated that changes were needed, then appropriate action would be taken. Panelists knew that new cutscores would not be computed from their recommendations, but they also knew that their recommendations would be taken seriously.

Most panelists indicated that they would not want to make further changes to the cutscores. Panelists recommended only 7 changes in the final cutscores for civics and 11 for writing. Six of the seven changes recommended in civics were at the Advanced level, and they were evenly split between raising and lowering the cutscore. Seven of the 11 changes recommended in writing were to lower the Advanced level cutscore.

Evaluations of Feedback Data in the NAEP ALS Process

Process evaluations are administered to panelists at the end of each day and at important transition points in the process. A total of seven process evaluations are administered during the five-day ALS process. The number and sequence of evaluations currently administered has been roughly the same since the 1994 process. Only four evaluations were administered in 1992 ALS cycle.

Many of the evaluation items have appeared on questionnaires throughout the eight years during which ACT has been NAGB's contractor for achievement levels. Some questionnaire items have been modified slightly during this period, some have been eliminated, and some have been added. The responses of panelists to the questionnaire items regarding the feedback will be reviewed in this section. Panelists have consistently been asked whether they understood the data, the source of the data, and how to use the data. They have consistently been asked whether they planned to use it, whether they used it, and how helpful it was to them during the rating process.

In addition to examining responses of panelists, observations will be shared regarding panelists' reactions to the feedback data.

While some general findings emerge regarding panelists' reactions to and evaluations of different types of feedback, it is clear that there is no unanimous favorite. The Reckase Charts were *very* well received by panelists in the 1998 ALS process, and they received very high ratings. Still, they were not the "favorite" of every panelist.

Cutscores and Standard Deviations

Panelists cannot directly "evaluate" the cutscores and standard deviations. Indeed, a major reason for setting achievement levels on NAEP is to give meaning to student performance that is intuitively more informative than a scale score. With no external frame of reference—no benchmarks for comparisons—panelists cannot judge whether cutscores are too high or too low.

ACT does not try to educate panelists to have a technically precise understanding of how cutscores are computed nor of the meaning of "standard deviations." Rather, we attempt to give them a general sense of the terms as indicators. No questions have been included on the process evaluations regarding

cutscores and standard deviations, per se. Some reactions by panelists are typical, and those can be shared here.

Mathematics, Reading, and the 1992 Writing NAEP were scaled on an across grade scale. Panelists understood that Basic level cutscores for fourth graders should be lower than those for eighth graders, and so forth. Whether the Advanced level cutscore for grade 8 should be higher, lower, or at about the same level as the Basic level cutscore for grade 12 was not known. Panelists seemed to identify some patterns among the cutscores as indicators of logical and reasonable outcomes. For example, finding that the difference between the cutscores for the Proficient and Advanced levels was less for grade 12 math than for grades 4 and 8 was seen as evidence that the cutscores were set about right. Panelists seem to expect less difference between the Proficient and Advanced cutscores than between those for the Basic and Proficient levels, for example.

Once panelists understand that the standard deviation is an indicator of their agreement or rater consistency, they pay attention to the relative size of the standard deviations associated with cutscores across grades. They cheer when the standard deviations decrease from one round to the next. They cheer when the standard deviations for their own grade level is lowest.

Student Performance Data
Panelists generally pay attention to student performance data, and they seem to find it both useful and interesting information. They discuss items for which performance seems particularly high or low. They frequently ask why student performance data for each achievement level are not provided to them.

Many questions about student performance data are included on process evaluation questionnaires. As the data in Table 1 show, panelists in each of the NAEP ALS processes generally respond positively to questions regarding their understanding of information about student performance data and how to use it. They tend to think that the amount of time devoted to instructions about the use of the data is *about right.* Their confidence in their ability to use the data for rating items is quite high. They generally indicate that the student performance data influenced their ratings *somewhat*, at least. Many panelists report that this is among *the most useful* feedback provided to them in the rating process.

Reckase Charts provide much more detailed data than student performance data. It does not seem likely that Reckase Charts will eliminate the usefulness of the student performance data, however. Although many panelists have requested student performance data, conditional with respect to the cutscores, panelists find the overall performance data helpful. Panelists value the single indicator of item difficulty. Some panelists have taken time to mark the p-value data on their Reckase Charts and to evaluate that relative to the expected performance of students at their cutscores and relative to the grade-level cutscores for the three achievement levels.

Interrater Consistency Data (rater location charts)
Panelists always seem to appreciate interrater consistency data. These data consistently appear to provide compelling evidence to panelists, although the compulsion takes different forms. Some panelists seem compelled to change their ratings so that they are more consistent with the other raters in the grade group while others seem compelled to be more different. Some panelists seem to take pride in being particularly low or high relative to the grade-level cutscore—they identify themselves as the "outliers;" others seem chagrined by this—they do not divulge this information. Panelists pay attention to these data.

The data in Table 2 report responses by panelists to evaluation questionnaire items regarding their understanding of interrater consistency data and their use of it in the NAEP ALS rating process. The format of data on the charts requires some explanation, but panelists seem to understand it quickly. They give very positive responses to items about the source, interpretation, and use of interrater consistency

data. They are able to interpret the data and to use it in rating items. They have high levels of confidence in their ability to use the data. The data in Table 2 provide ample evidence that this feedback is understood and is perceived to be useful. Nonetheless, responses do not indicate that the data strongly influence panelists' ratings and the cutscores that they set.

Whole Booklet Data

The whole booklet data were added in 1994, and they have been used in all NAEP ALS processes since then. As explained earlier, two types of whole booklet data are provided: booklets scored at the cutpoints are reviewed by panelists and the proportion of total possible points required for performance at the cutscores is reported to panelists.

Panelists have generally been positive in their evaluations of these data. Some panelists have commented that they would like to have the whole booklets to review after Round 2, as well as after Round 1. Some find that this exercise discloses very valuable information about student performance relative to the cutscores. Other panelists have commented that these data are not very helpful because the individual items are not scored; because a NAEP booklet contains relatively few items—not enough to provide evidence of performance at the borderline; and because students can earn high scores without answering easy items and without answering most or all items.

Panelists find the whole booklet feedback reports useful, although they show little enthusiasm for these reports. Some panelists seem to convert the "percentage of total possible points" to a letter-grade scale. If, for example, fewer than 50% of total possible points on a test form are required for performance at the Basic cutscore, panelists often complain that this is too low because they consider such performance to be "failing." Similarly, many panelists seem to believe that at least 90% of the total possible points are needed to consider performance as Advanced.

The data in Table 3 show panelists' responses to evaluation questionnaire items regarding the whole booklet data. Panelists report relatively high levels of understanding, clarity, confidence, and so forth regarding the data. The data are less well understood than those previously reviewed, however. Relative to other feedback, the whole booklet data generally get lower ratings. This is an aspect of the process requiring more attention. Logically, the holistic data are likely to be more difficult for panelists to incorporate into the item-by-item rating process than item-level data. In any case, more attention to instructions regarding the source of the data, the interpretation of the data, and how to use the data seems appropriate.

Intrarater Consistency Data and Reckase Charts

As explained earlier, the provision of intrarater consistency feedback data has been a challenge. Finding a way to present the data has been challenging, as has finding a way to instruct panelists in the meaning and use of the data. The data were dropped from the feedback provided in the 1996 Science NAEP ALS process, and ACT gave considerable thought to improving the feedback.

Evaluations by panelists do not reveal this negative impression regarding intrarater consistency feedback data. Responses reported in Table 4.1 indicate that the panelists understood the information regarding what the data were, how to interpret them, and how to use them in rating items. They also expressed high levels of confidence in their ability to use the data.

Analyses of rating data did not support this confidence. Panelists did not seem to use the data to increase the consistency of their ratings. Panelists with higher levels of inconsistency in their item ratings did not make larger adjustments to ratings or more adjustments to ratings than other panelists. Instructions to panelists regarding feedback always stress that the feedback data are information to be used or not used; panelists are to make that decision. It seemed unlikely, however, that panelists with extremely low levels

of intrarater consistency were generally rejecting the information provided by this feedback and deciding to keep "inconsistent" ratings for all or most of the items on the lists. It seemed more likely that they simply could not understand the data and how to use the reports. Facilitators generally felt low confidence in their ability to explain the data to panelists and to provide intuitive explanations regarding the interpretation and use of the intrarater consistency data. After reviewing the data with TACSS, ACT decided to drop this feedback from the process.

Starting in 1998, Reckase Charts were included as a new type of feedback. The Reckase Charts are discussed under "intrarater consistency" feedback although they far exceed the type of intrarater consistency feedback provided prior to1998.

Reckase Charts were new, and many questions were included on the evaluation questionnaires to collect panelists' reactions to various aspects of the Reckase Charts. Table 4.2 contains panelists' evaluations of Reckase Charts. As the large number of items indicates, many types of rating consistency were evaluated with the Reckase Charts. Panelists seemed to understand the information about the data in the charts, how to interpret them, and how to use them in item ratings. They were generally quite confident in their ability to use the data. And, many panelists reported that they used the charts to adjust their ratings. Further, many panelists reported that the data influenced their ratings.

Panelists were enthusiastic in their positive evaluations of the Reckase Charts. They wrote positive comments about the ease of using them and the richness of data provided by the charts.

Consequences Data

The final type of feedback data provided to panelists is consequences data. These data have been provided since 1994, but they were only provided during the ratings for the 1998 ALS NAEP process. The impact of consequences data on ratings was examined during field trials for the 1998 ALS process, as well as during the pilot study for each subject. Results from those studies were confirmed by the ALS findings. Namely, consequences data do not have a large, measurable impact on the cutscores set by panelists.

The data in Table 5 show responses by panelists to evaluation questionnaire items regarding their reactions to and uses of consequences data. After instructions on their use, panelists rated items in the third item-by-item round of ratings. Panelists seemed to understand the instructions on the source of the data, the use of the data, and how to interpret the data. They felt confident that they could use the data. Yet, the consequences data were perceived to have relatively little impact on their ratings in Round 3, and panelists gave relatively low ratings to the usefulness of the consequences data for their Round 3 ratings.

Overall Evaluations of Different Feedback

Panelists' evaluations of the feedback in the final process evaluation questionnaires are reported in Tables 6-8. The data were divided among the various ALS cycles because the formats of the questions changed somewhat from one period to the next. For the first period, reported in Table 6, panelists were asked to rank different information according to how helpful it was in the process of rating items. Training in the 1992 process included an exercise during which panelists selected items that about 80% of students performing at the borderline of an achievement level should answer correctly. That will not be included in this discussion since it is not feedback information. Among the information included in that list were the achievement levels descriptions, and they will be discussed. Although they are not feedback, the ALDs are so integral to the process that they must be included in this discussion of the relative impact of feedback information for the rating process.

As can be seen from the data in Table 6, panelists generally rated the ALDs as ***the*** most helpful information to them in the item rating process. Our goal is to have the ALDs be the most important

impact on the item ratings. Student performance data were the second most helpful information to panelists in math and reading, and intrajudge consistency information was ranked second by the 1992 Writing ALS panelists.

Table 7 presents the same data for the second cycle including geography, history, and science. Note that the evaluation of whole booklet feedback was inadvertently omitted from the final process evaluation questionnaire for science. Achievement levels descriptions were again rated as the most helpful information for panelists to use in the item rating process. Of the four types of feedback included on the questionnaires for geography and U.S. history, panelists in both subjects gave the lowest rating to whole booklet feedback. Geography panelists rated interjudge and intrajudge consistency feedback equally high, and these types of feedback tied as second to the ALDs in terms of helping panelists in the rating process. History panelists rated intrajudge consistency feedback as second, student performance data as third, and interjudge consistency feedback as fourth.

Data in Table 8 are for civics and writing, subjects in the 1998 NAEP ALS process. Note that Reckase Charts have replaced the ALDs as *the* most helpful information for rating items. While we are gratified that panelists find the Reckase Charts helpful, we are concerned that they are having too great an impact on panelists' ratings. The ALDs *must be the guide* to setting achievement levels. We do not have data to help determine just how to interpret this finding. Panelists reported that Reckase Charts influenced their ratings more often or to a greater extent than any other form of feedback included in this 1998 ALS process. We cannot judge whether that means that the influence of Reckase Charts had an impact on ratings that was as great or greater than the ALDs. Unfortunately, we did not think to include the ALDs on the list of influences on ratings in the questionnaire. When asked to judge the *helpfulness*, however, Reckase Charts replaced ALDs in the process.

As a final indication of the role of feedback in the ALS process, it will perhaps be helpful to examine the data in Table 9 reporting responses by panelists in all 8 ALS processes conducted for NAGB by ACT. All panelists have been asked to evaluate the extent to which the ALS process has allowed them the opportunity to use their best judgment, the extent to which the outcomes of the process are defensible, the extent to which the outcomes will be judged to be reasonable, and whether they would be willing to sign a statement recommending adoption of the achievement levels.

The results presented in Table 9 do not reflect an expected pattern of responses. Many changes have been made to improve the NAEP ALS process. Members of the TACSS were positively impressed by the analyses of the 1998 process. All indications were that this was the best NAEP ALS process to date. Given that the process has been in a steady state of improvement, one would not expect to find that panelists gave the highest ratings on these items in the first cycle. Only the 1998 NAEP Writing panelists gave a higher rating to the process in terms of the opportunity to use their best judgment. Interestingly, writing panelists in the 1992 and 1998 process gave the highest ratings to the defensibility of outcomes of the process. The 1998 writing panelists gave relatively low ratings to the reasonableness of the outcomes. They had seen consequences data and their ratings were informed by the data; the 1992 panelists had no information about the consequences of the outcomes they judged as reasonable. The 1998 Writing ALS panelists had quite a conversation in each grade group regarding the consequences data. They strongly urged that all reports of student performance on the 1998 Writing NAEP include an explanation about the circumstances of assessing students such as the 25-minute time limit, the fact that students have no opportunity to prepare in advance for the topic on which their writing is assessed, and so forth.

U.S. history panelists gave the lowest ratings to all of these attributes of the process. These panelists were informed about the consequences of their ratings after the final round of ratings—before responding to this questionnaire. The performance of students at or above the Advanced level was low for all three grade levels, and only slightly over half of the grade 12 students performed at or above the Basic level.

Despite these findings, however, panelists did not recommend changes in cutscores after seeing consequences data and 80% of the panelists would *probably* or *definitely* sign a statement recommending the achievement levels.

In previous analyses of panelists' responses, civics panelists were found to have a generally low propensity toward positive responses. The group was judged, as a whole, to be the most contentious group. Civics panelists seemed suspicious of the process. There were small groups of panelists in each of the grade groups who seemed intent on fomenting distrust.

The same process implemented for civics was implemented for writing. The positive evaluations by writing panelists were unusually high, and those by the civics panelists were unusually low. Compared to geography and U.S. history panelists (the other social sciences), relatively high percentages of civics panelists responded *to a great extent* when asked to rate the process with respect to the opportunity to use their best judgment, and the defensibility and reasonableness of the outcomes. Despite these relative high levels of the most positive response, however, relatively small percentages gave the two highest evaluations to these items. This means that their overall or average evaluations were relative low.

Summary

Standard setting is a judgmental process. It is important that the judgments be made by people who are well trained to do the task and well informed about their performance of the task. Feedback to panelists about their ratings is one of the most important aspects of the ACT/NAGB ALS process. The NAEP ALS process includes a variety of feedback to panelists throughout the rating process. The reactions of panelists to the feedback is monitored throughout the process by a series of questionnaires administered to panelists to record their opinions about a comprehensive set of attributes associated with the feedback. Panelists' responses indicate high levels of understanding about and confidence in using the feedback. Panelists' responses do not always correspond to on-site observations and subsequent analyses of ratings. The Reckase Charts appear to be a significant addition to the array of feedback information provided to panelists. Further analyses must be conducted to determine whether the Reckase Charts supplant the primary role of achievement levels descriptions in the item rating process.

References

ACT (1993a). *Description of mathematics achievement levels process and proposed achievement levels descriptions. National Assessment of Educational Progress in mathematics: Final report.* Iowa City, IA: Author.

ACT (1993b). *Setting achievement levels on the 1992 National Assessment of Educational Progress in reading: Final report.* Iowa City, IA: Author.

ACT (1993c). *Setting achievement levels on the 1992 National Assessment of Educational Progress in writing: Final report.* Iowa City, IA: Author.

ACT (1994a). *Results of the 1994 U.S. History National Assessment of Educational Progress achievement levels-setting pilot study.* Iowa City, IA: Author.

ACT (1994b). *Results of the 1994 Geography National Assessment of Educational Progress achievement levels-setting pilot study.* Iowa City, IA: Author.

ACT (1995). *Preliminary report on the 1994 National Assessment of Educational Progress achievement levels-setting process for U.S. history, and geography.* Iowa City, IA: Author.

ACT (1997a). *Setting achievement levels on the 1996 National Assessment of Educational Progress in science. Final report, Executive summary.* Iowa City, IA: Author.

ACT (1997b). *Setting achievement levels on the 1996 National Assessment of Educational Progress in science. Final report, Volume 1. Pilot study 1.* Iowa City, IA: Author.

ACT (1997c). *Setting achievement levels on the 1996 National Assessment of Educational Progress in science. Final report, Volume II. Pilot study 2.* Iowa City, IA: Author.

ACT (1997d). *Setting achievement levels on the 1996 National Assessment of Educational Progress in science. Final report, Volume III. Achievement levels-setting study.* Iowa City, IA: Author.

ACT (1997e). *Setting achievement levels on the 1996 National Assessment of Educational Progress in science: Final report, Volume IV. Validity evidence special studies.* Iowa City, IA: Author.

ACT (1997f). *Setting achievement levels on the 1996 National Assessment of Educational Progress in science. Final report, Volume V. Technical decisions NAGB actions.* Iowa City, IA: Author.

Hanick, P.L. (1999a). *1998 Civics NAEP Achievement Levels-Setting meeting: Summary report of process evaluation questionnaires.* A paper presented at the meeting of the Technical Advisory Committee for Standard Setting (TACSS), Atlanta.

Hanick, P.L. (1999b). *1998 Writing NAEP Achievement Levels-Setting meeting: Summary report of process evaluation questionnaires*. A paper presented at the meeting of the Technical Advisory Committee for Standard Setting (TACSS), Atlanta.

Hanick, P.L. (1999c). *Follow-up analyses to irregular responses from Civics ALS evaluation questionnaires.* A paper presented at the meeting of the Technical Advisory Committee for Standard Setting (TACSS), Atlanta.

Loomis, S.C. (1999a). *NAEP 1998 Writing ALS Summary.* A paper presented at the meeting of the Technical Advisory Committee for Standard Setting (TACSS), Atlanta.

Loomis, S.C. (1999b). *Panelists' reactions to consequences data: Civics ALS.* A paper presented at the meeting of the Technical Advisory Committee for Standard Setting (TACSS), Atlanta.

Loomis, S.C. (1999c). *Panelists' reactions to consequences data: Writing ALS.* A paper presented at the meeting of the Technical Advisory Committee for Standard Setting (TACSS), Atlanta.

Loomis, S.C., Bay, L., Yang, W.L., & Hanick, P.L. (1999). *Field trials to determine which rating method(s) to use in the 1998 NAEP achievement levels-setting process.* Paper presented at the meeting of the NCME, Montreal.

Luecht, R.M. (1992, September). Some applications of item response theory to standard setting. In Appendix N of ACT, *Description of mathematics achievement levels-setting process and proposed achievement level descriptions (Volume II).* Iowa City, IA: ACT.

National Academy of Education (1993). *Setting Performance Standards for Student Achievement*, Robert Glaser, Robert Linn, and George Bohrnstedt, eds. Panel on the Evaluation of the NAEP Trial State Assessment. Stanford, CA: Author.

Reckase, M.D. & Bay, L. (1999). *Comparing two methods for collecting test-based judgments.* Paper presented at the annual meeting of the National Council on Measurement in Education, 1999, Montreal.

Table 1
Evaluating Student Performance Data: Mean Responses

Question	1992 Math	1992 Reading	1992 Writing	1994 Geography	1994 US History	1996 Science	1998 Civics	1998 Writing
Information about the source of student performance data to be considered during the second rating session was: (5=Absolutely Clear; 1=Not at all Clear)	4.24	4.25	4.14	4.36	4.17	4.41	4.03	4.42
Instructions on the interpretation of student performance data to be considered during the second item-rating session were: (5=Absolutely Clear; 1=Not at all Clear)	4.32	4.22	4.00	4.32	4.11	4.39	4.02	4.23
Instructions on the use of student performance data to be considered during the second item-rating session were: (5=Absolutely Clear; 1=Not at all Clear)	4.35	4.15	3.98	4.28	4.13	4.39	4.02	4.14
If the length of time spent on instructions concerning student performance data were to be changed, I would recommend: (5=Far More Time; 1=Far Less Time)	2.88	3.10	3.33	3.00	3.07	3.12	3.36	3.30
Following instructions, the most accurate description of my level of confidence in my ability to use student performance data during the second item-rating session is that I was: (5=Totally Confident; 1=Not at all Confident)	4.03	4.00	3.87	4.11	4.00	4.08	4.02	4.03
When I provided ratings during the second item-rating session, my judgments were influenced by student performance data: (5=Greatly; 1=Not at All)	3.41	3.51	3.61	3.55	3.59	3.27	3.85	3.69
The most useful information was the P-value data, i.e., % of students who correctly answered or the average score on each item. (Second session) (5=Totally Agree; 1=Totally Disagree)	N/A	N/A	N/A	3.64	3.68	3.26	3.50	3.48
I used the P-value data to adjust my ratings during Round 2. (5=To a Great Extent; 1=Not at All)	N/A	N/A	N/A	N/A	N/A	N/A	3.61	3.66
I plan to use only the student performance data to adjust my ratings for Round 2. (5=Totally Agree; 1=Totally Disagree)	N/A	N/A	N/A	N/A	N/A	N/A	1.90	1.80
When I provided ratings during the third item-rating session, my judgments were influenced by student performance data: (5=Greatly; 1=Not at All)	2.69	2.88	3.28	N/A	N/A	N/A	3.48	3.31
The most useful information was the "P-value Feedback" (% of students who correctly answered or the average score) on each item. (Third session) (5=Totally Agree; 1=Totally Disagree)	N/A	N/A	N/A	3.30	3.28	3.28	3.39	3.38
I used the P-value data to adjust my ratings during Round 3. (5=To a Great Extent; 1=Not at All)	N/A	N/A	N/A	N/A	N/A	2.97	3.31	3.03

Table 2
Evaluating Interrater Consistency: Mean Responses

Question	1992 Math	1992 Reading	1992 Writing	1994 Geography	1994 US History	1996 Science	1998 Civics	1998 Writing
Information about the source of data on my ratings relative to those of other panelists to be considered during the second item-rating session was: (5=Absolutely Clear; 1=Not at all Clear)	4.25	4.02	4.00	4.48	4.26	4.52	4.19	4.55
Instructions on the interpretation of data on my ratings relative to those of other panelists to be considered during the second item-rating session were: (5=Absolutely Clear; 1=Not at all Clear)	4.30	4.10	3.97	4.39	4.23	4.55	4.21	4.53
Instructions on the use of data on my ratings relative to those of other panelists to be considered during the second item-rating session were: (5=Absolutely Clear; 1=Not at all Clear)	4.22	4.07	3.95	4.32	4.13	4.50	4.13	4.43
If the length of time spent on instructions concerning data on my ratings relative to those of other panelists were to be changed, I would recommend: (5=Far More Time; 1=Far Less Time)	2.94	3.22	3.23	2.95	3.02	3.03	3.29	3.24
Following instructions, the most accurate description of my level of confidence in my ability to use data on my ratings relative to those of other panelists during the second item-rating session is that I was: (5=Totally Confident; 1=Not at all Confident)	4.12	4.08	3.95	4.19	4.10	4.18	4.14	4.38
When I provided ratings during the second item-rating session, my judgments were influenced by data on my ratings relative to those of other panelists: (5=Greatly; 1=Not at All)	2.31	2.75	2.91	3.40	3.04	3.16	3.55	3.73
The most useful information was the report on my average ratings/cutpoints at each achievement level relative to others in my group (interrater consistency). (Second session) (5=Totally Agree; 1=Totally Disagree)	N/A	N/A	N/A	3.52	3.35	3.10	3.24	3.57
I used the rater location data to adjust my ratings during Round 2. (5=To a Great Extent; 1=Not at All)	N/A	N/A	N/A	N/A	N/A	3.04	3.23	3.35
I plan to use only the student performance data to adjust my ratings for Round 2. (5=Totally Agree; 1=Totally Disagree)	N/A	N/A	N/A	N/A	N/A	N/A	2.01	1.73
When I provided ratings during the third item-rating session, my judgments were influenced by data on my ratings relative to those of other panelists: (5=Greatly; 1=Not at All)	2.43	2.61	2.88	N/A	N/A	N/A	3.42	3.64
The most useful information was the interrater consistency data (report on my average ratings/cutpoint at each achievement level relative to others in my group). (Third session) (5=Totally Agree; 1=Totally Disagree)	N/A	N/A	N/A	3.70	3.56	3.97	3.42	3.58
I used the rater location data to adjust my ratings during Round 3. (5=To a Great Extent; 1=Not at All)	N/A	N/A	N/A	N/A	N/A	2.98	3.21	3.19

Table 3
Evaluating Whole Booklet Feedback Data: Mean Responses

Question	1992 Math	1992 Reading	1992 Writing	1994 Geography	1994 US History	1996 Science	1998 Civics	1998 Writing
Instructions on the interpretation of student performance exhibited via the Whole Booklet Exercise were: (5=Absolutely Clear; 1=Not at all Clear)	N/A	N/A	N/A	4.20	4.04	4.33	N/A	N/A
Instructions on the use of information from the Whole Booklet Exercise and Feedback were: (5=Absolutely Clear; 1=Not at all Clear)	N/A	N/A	N/A	4.19	4.06	4.28	N/A	N/A
If the length of time spent on instructions about the Whole Booklet Exercise and Feedback were to be changed, I would recommend: (5=Far More Time; 1=Far Less Time)	N/A	N/A	N/A	3.12	2.96	3.16	N/A	N/A
The most accurate description of my level of confidence in my ability to use data about student performance based on the Whole Booklet Exercise and Feedback session is that I was: (5=Totally Confident; 1=Not at all Confident)	N/A	N/A	N/A	4.11	3.99	4.00	N/A	N/A
When I provided ratings during the second rating session, my judgments were influenced by the Whole Booklet Exercise and Feedback: (5=Greatly; 1=Not at All)	N/A	N/A	N/A	3.26	3.35	3.03	3.36	3.52
The most useful information was the report on student performance via the Whole Booklet Exercise and Feedback. (Second session) (5=Totally Agree; 1=Totally Disagree)	N/A	N/A	N/A	3.19	3.23	2.94	2.84	2.91
I used data from the Whole Booklet Exercise and Feedback to adjust my ratings during Round 2. (5=To a Great Extent; 1=Not at All)	N/A	N/A	N/A	N/A	N/A	2.83	2.84	2.84
I plan to use only data from the Whole Booklet Exercise and Feedback to adjust my ratings for Round 2. (5=Totally Agree; 1=Totally Disagree)	N/A	N/A	N/A	N/A	N/A	N/A	2.12	1.91
When I provided ratings during the third item-rating session, my judgments were influenced by the Whole Booklet Exercise and Feedback: (5=Greatly; 1=Not at All)	N/A	N/A	N/A	N/A	N/A	N/A	2.84	2.97
The most useful information was the report on student performance via the Whole Booklet Exercise and Feedback. (Third session) (5=Totally Agree; 1=Totally Disagree)	N/A	N/A	N/A	3.18	3.06	N/A	2.75	2.64
I used the Whole Booklet Exercise and Feedback to adjust my ratings during Round 3. (5=To a Great Extent; 1=Not at All)	N/A	N/A	N/A	N/A	N/A	2.59	2.65	2.33

Table 4.1
Evaluating Intrarater Consistency Data: Mean Responses

Question	1992 Math	1992 Reading	1992 Writing	1994 Geography	1994 US History	1996 Science	1998 Civics	1998 Writing
Information about the source of data about the consistency of my ratings to be considered during the third rating session was: (5=Absolutely Clear; 1=Not at all Clear)	4.12	4.32	4.17	N/A	N/A	N/A	N/A	N/A
Instructions on the interpretation of data about the consistency of my ratings to be considered during the third item-rating session were: (5=Absolutely Clear; 1=Not at all Clear)	4.25	4.24	4.12	4.23	4.26	N/A	N/A	N/A
Instructions on the use of data about the consistency of my ratings to be considered during the third item-rating session were: (5=Absolutely Clear; 1=Not at all Clear)	4.25	4.31	4.18	4.26	4.35	N/A	N/A	N/A
If the length of time spent on data about the consistency of my ratings were to be changed, I would recommend: (5=Far More Time; 1=Far Less Time)	2.99	3.05	3.17	2.89	2.90	N/A	N/A	N/A
Following instructions, the most accurate description of my level of confidence in my ability to use data about the consistency of my ratings I was: (5=Totally Confident; 1=Not at all Confident)	4.31	4.24	4.21	4.14	4.23	N/A	N/A	N/A
When I provided ratings during the third item-rating session, my judgments were influenced by data about the consistency of my ratings: (5=Greatly; 1=Not at All)	3.21	3.27	3.94	N/A	N/A	N/A	N/A	N/A
The most useful information was the data about the consistency of my ratings. (Third session) (5=Totally Agree; 1=Totally Disagree)	N/A	N/A	N/A	3.55	3.99	N/A	N/A	N/A

Table 4.2
Evaluating Reckase Charts: Mean Responses

Question	1998 Civics	1998 Writing
Information about the source of data in the Reckase Charts for each item to be considered during the second rating session was: (5=Absolutely Clear; 1=Not at all Clear)	3.96	4.41
Instructions on the interpretation of data in the Reckase Charts for each item to be considered during the second rating session were: (5=Absolutely Clear; 1=Not at all Clear)	3.96	4.48
Instructions on the use of data in the Reckase Charts for each item to be considered were: (5=Absolutely Clear; 1=Not at all Clear)	4.10	4.43
If the length of time spent on instructions concerning Reckase Charts for each item were to be changed, I would recommend: (5=Far More Time; 1=Far Less Time)	3.29	3.20
When I provided ratings during the second rating session, my judgments were influenced by the data in the Reckase Charts. (5=Greatly; 1=Not at All)	3.98	4.38
The most useful information was the Reckase Charts. (Second session) (5=Totally Agree; 1=Totally Disagree)	3.77	4.16
I used the Reckase Charts to adjust my ratings during Round 2. (5=To a Great Extent; 1=Not at All)	3.85	4.30
The Reckase Charts helped me adjust my ratings to better reflect my concept of Borderline Basic performance. (5=Totally Agree; 1=Totally Disagree)	3.93	4.06
The Reckase Charts helped me adjust my ratings to better reflect my concept of Borderline Proficient performance. (5=Totally Agree; 1=Totally Disagree)	3.81	4.07
The Reckase Charts helped me adjust my ratings to better reflect my concept of Borderline Advanced performance. (5=Totally Agree; 1=Totally Disagree)	3.83	4.02
When I provided ratings during the third rating session, my judgments were influenced by the Reckase Charts. (5=Greatly; 1=Not at All)	3.79	4.13
The most useful information was from the Reckase Charts. (Third session) (5=Totally Agree; 1=Totally Disagree) .	3.80	4.16
I used the Reckase Charts data to adjust my ratings during Round 3. (5=To a Great Extent; 1=Not at All)	3.84	4.08

Question	1998 Civics	1998 Writing
Of the information available to me during the rating process, the Reckase Charts (student performance at each scale score) were most helpful. (5=Totally Agree; 1=Totally Disagree)	4.24	4.42
I would characterize the ease of reading the Reckase Charts as: (5=Very Easy; 1=Not at all Easy)	4.09	4.82
I would characterize the ease of marking the Reckase Charts as: (5=Very Easy; 1=Not at all Easy)	4.05	4.88
My level of understanding how the ratings for items displayed on the Reckase Charts are related to my own cutscores was: (5=Totally Adequate; 1=Not at all Adequate)	4.56	4.68
My level of understanding how the ratings for items displayed on the Reckase Charts are related to my grade group cutscores was: (5=Totally Adequate; 1=Not at all Adequate)	4.59	4.73

Table 5
Evaluating Consequences Data: Mean Responses

Question	1998 Civics	1998 Writing
Information about the source of consequences data reporting the % of scores at or above each score point to be considered during the third rating session was: (5=Absolutely Clear; 1=Not at all Clear)	4.47	N/A
Instructions on the interpretation of consequences data reporting the % of scores at or above each score point to be considered during the third rating session were: (5=Absolutely Clear; 1=Not at all Clear)	4.46	N/A
Instructions on the use of consequences data reporting the % of scores at or above each score point to be considered during the third rating session were: (5=Absolutely Clear; 1=Not at all Clear)	4.37	N/A
If the length of time spent on instructions concerning consequences data reporting the % of scores at or above each score point were to be changed, I would recommend: (5=Far More Time; 1=Far Less Time)	3.00	N/A
Following instructions, the most accurate description of my level of confidence in my ability to use consequences data reporting the % of scores at or above each score point during the third rating session is that I was: (5=Totally Confident; 1=Not at all Confident)	4.34	N/A
When I provided ratings during the third rating session, my judgments were influenced by consequences data. (5=Greatly; 1=Not at All)	3.07	3.34
The most useful information was the consequences data. (Third session) (5=Totally Agree; 1=Totally Disagree)	2.87	3.13
I used the consequences data to adjust my ratings during Round 3. (5=To a Great Extent; 1=Not at All)	2.94	3.05
Of the information available to me during the rating process, the consequences data (% of scores at or above each cutpoint) was most helpful. (5=Totally Agree; 1=Totally Disagree)	3.37	3.75

Table 6

Rating Most Helpful Feedback in the 1992 NAEP ALS Cycle

	Math 1992	Reading 1992	Writing 1992
Student Performance Data			
%#1	16.18	15.00	1.52
% #1 + #2	54.4	38.33	19.70
Mean (1=Most; 5=Least Helpful)	2.51	2.67	3.75
Intrajudge Consistency			
%#1	5.88	6.67	9.09
% #1 + #2	22.06	13.34	37.88
Mean (1=Most; 5=Least Helpful)	3.38	3.52	2.76
Interjudge Consistency			
%#1		8.33	4.55
% #1 + #2		21.66	12.13
Mean (1=Most; 5=Least Helpful)		3.65	3.83
Selection of Illustrative Items with RP=.8			
%#1	2.94	11.67	3.03
% #1 + #2	22.06	38.34	31.82
Mean (1=Most; 5=Least Helpful)	3.71	3.22	3.19
Achievement Levels Descriptions			
%#1	73.53	50.00	72.73
% #1 + #2	86.77	71.67	80.31
Mean (1=Most; 5=Least Helpful)	1.47	1.91	1.42

Question: Please rank the following list in terms of the helpfulness to you in rating items. Use a *1* for the *most helpful*, *2* for the *second most helpful* and so forth—so that *5* represents the *least helpful*.

Table 7
Agreement with Statements Regarding the Most Helpful Feedback in the 1994-96 NAEP ALS Cycle

	Geography 1994	History 1994	Science 1996
Student Performance Data			
%#1	21.4	29.8	14.9
% #1 + #2	69.0	72.7	56.4
Mean (1=Totally Agree; 5=Totally Disagree)	3.88	4.01	3.58
Intrarater Consistency			
%#1	27.4	31.0	
% #1 + #2	71.4	78.6	
Mean (1=Totally Agree; 5=Totally Disagree)	3.90	4.13	
Interrater Consistency			
%#1	27.4	25.0	14.9
% #1 + #2	72.6	70.2	51.1
Mean (1=Totally Agree; 5=Totally Disagree)	3.98	3.95	3.45
Whole Booklet Exercise/Feedback			
%#1	21.4	21.4	
% #1 + #2	71.4	55.9	
Mean (1=Totally Agree; 5=Totally Disagree)	3.88	3.68	
Achievement Levels Descriptions			
%#1	44.0	41.7	21.3
% #1 + #2	84.5	75.0	73.4
Mean (1=Totally Agree; 5=Totally Disagree)	4.24	4.19	3.93

Question: X (student performance data, interrater consistency, etc.) was the most helpful.

Table 8
Agreement with Statements Regarding the Most Helpful Feedback in the 1998 NAEP ALS Cycle

	Civics 1998	Writing 1998
Student Performance Data		
%#1	37.9	20.5
% #1 + #2	66.6	60.3
Mean (1=Totally Agree; 5=Totally Disagree)	4.04	3.72
Reckase Charts		
%#1	48.3	58.0
% #1 + #2	78.2	85.3
Mean (1=Totally Agree; 5=Totally Disagree)	4.24	4.42
Rater Location Data		
%#1	31.0	27.3
% #1 + #2	52.8	69.3
Mean (1=Totally Agree; 5=Totally Disagree)	3.75	3.92
Whole Booklet Exercise		
%#1	24.1	15.9
% #1 + #2	41.3	53.4
Mean (1=Totally Agree; 5=Totally Disagree)	3.48	3.51
Whole Booklet Feedback		
%#1	17.2	13.6
% #1 + #2	36.7	54.5
Mean (1=Totally Agree; 5=Totally Disagree)	3.37	3.52
Achievement Levels Descriptions		
%#1	39.1	39.8
% #1 + #2	64.4	79.6
Mean (1=Totally Agree; 5=Totally Disagree)	4.00	4.14

Question: Of the information available to me during the rating process, X (student performance data, rater location data, etc.) was the most helpful.

Table 9
Rating the Overall ALS Process

	Math 1992	Reading 1992	Writing 1992	Geography 1994	History 1994	Science 1996	Civics 1998	Writing 1998
I feel that this process provided me an opportunity to use my best judgment in rating items to set achievement levels for the NAEP assessment.								
% Great Extent =5	57.35	53.33	54.55	38.1	32.1	44.7	47.1	61.4
% 4 + 5	88.23	85.00	87.88	86.9	80.9	92.6	71.2	92.1
Mean	4.46	4.38	4.41	4.27	4.19	4.36	4.14	4.53
(5=To a Great Extent; 1=Not at All)								
I feel that this process has produced achievement levels that are defensible.								
% Great Extent =5	42.65	38.33	46.97	31.0	25.0	35.1	42.5	46.0
% 4 + 5	85.30	83.33	87.88	84.6	69.0	75.5	72.4	83.9
Mean	4.25	4.19	4.32	4.14	3.90	4.02	4.14	4.26
(5=To a Great Extent; 1=Not at All)								
I feel that this process has produced achievement levels that will generally be considered reasonable.								
% Great Extent =5	50.00	43.33	59.09	38.1	19.0	30.9	46.0	39.8
% 4 + 5	88.24	90.00	92.42	82.1	63.0	70.3	75.9	86.4
Mean	4.39	4.33	4.54	4.23	3.75	3.88	4.19	4.26
(5=To a Great Extent; 1=Not at All)								
I would be willing to sign a statement (after reading it, of course) recommending the use of the achievement levels resulting from this achievement levels-setting procedure.								
% Definitely	55.88	65.00	62.12	57.1	33.3	48.9		64.7
% Probably	42.65	28.33	34.85	36.9	47.6	43.6		34.1
(4=Definitely; 3=Probably; 2=Probably Not; 1=Definitely Not)								

Figure 1
Cutscores and Standard Deviations: 1992 Math ALS

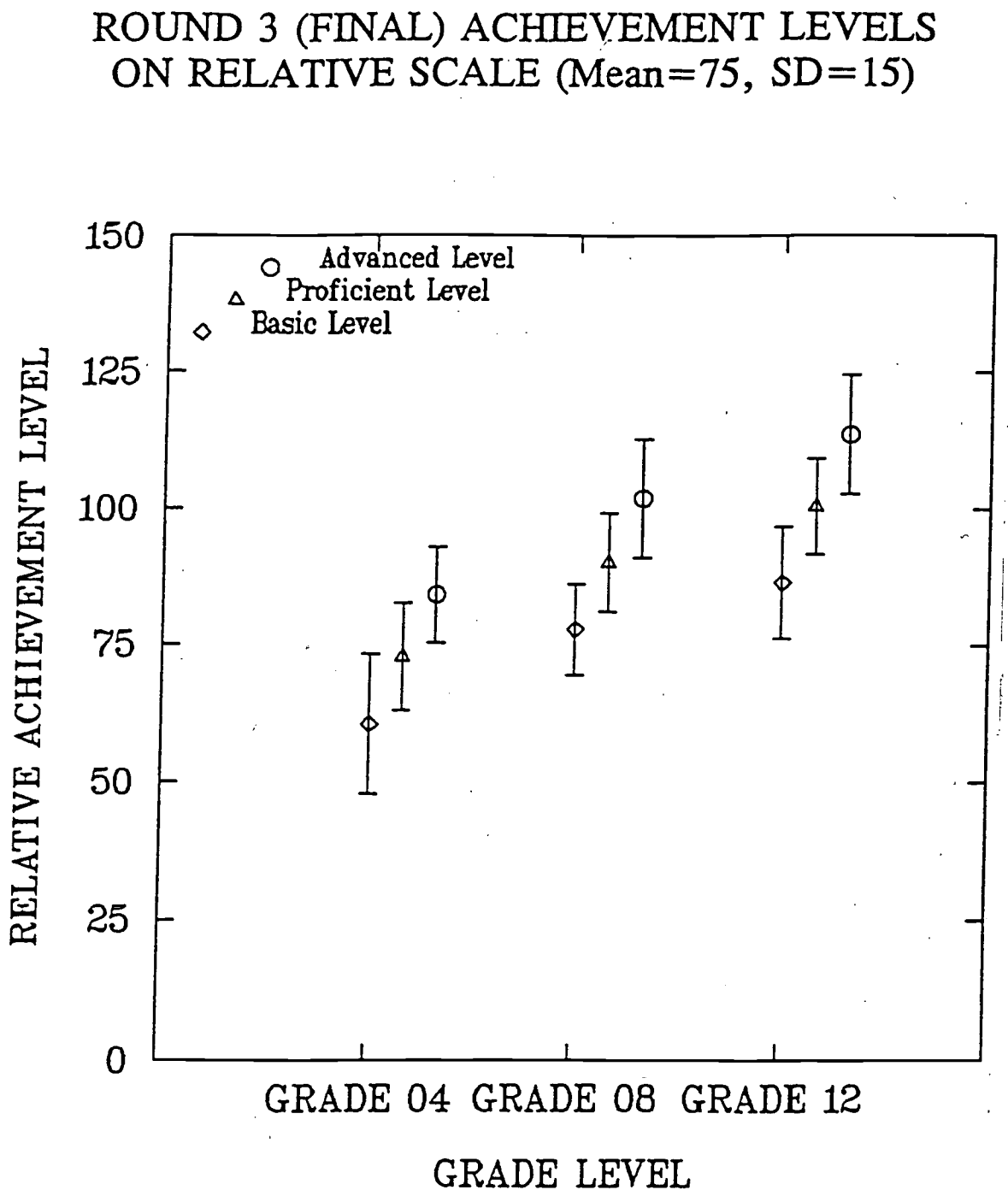
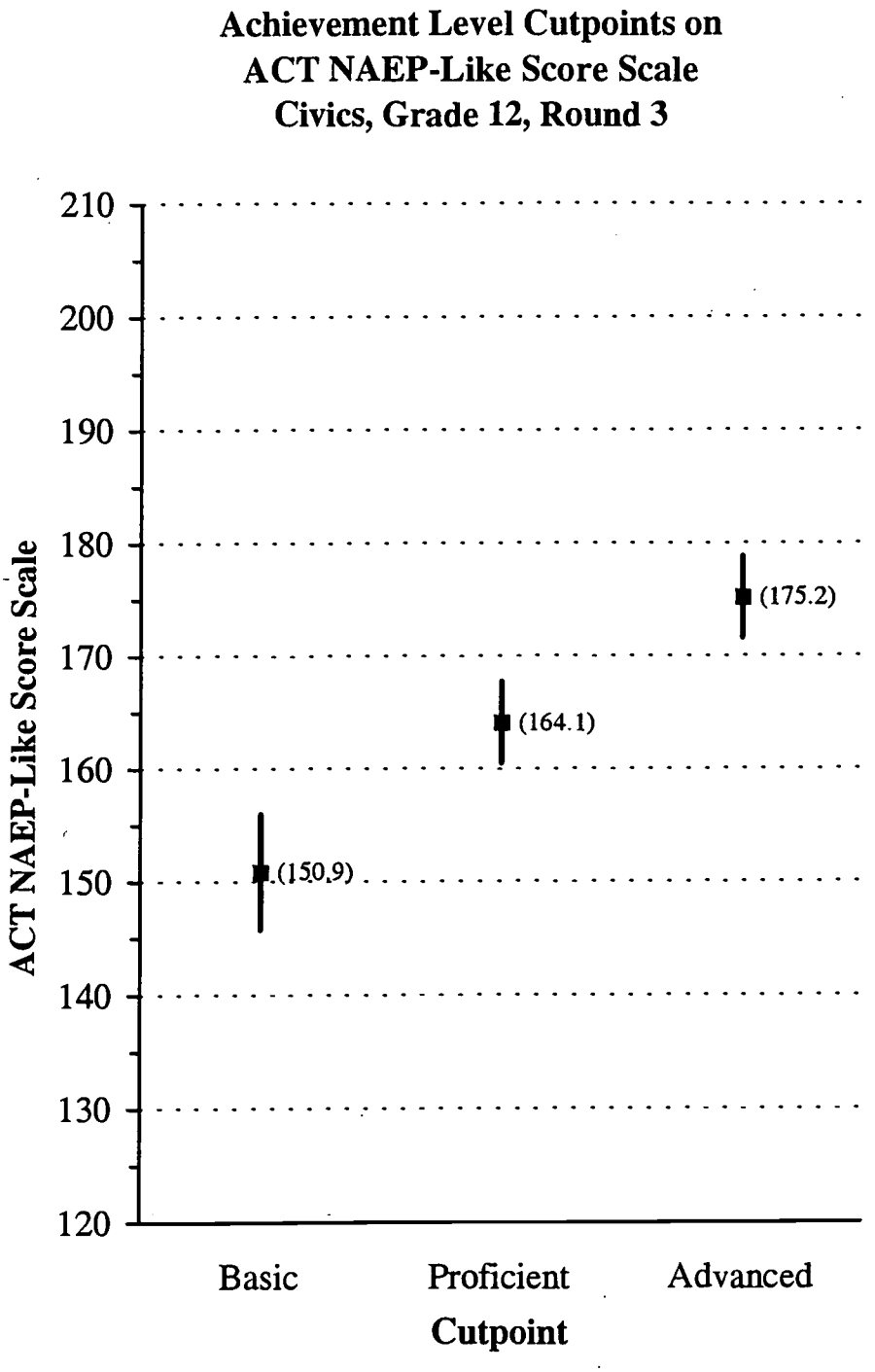


Figure 2
Cutscores and Standard Deviations: 1998 Civics



Note. A solid square represents the cutpoint and the bar represents variability in cutpoints among panelists.

Figure 3
Student Performance Data: 1992 Math ALS

Item No.	Number of Students Out of 100
Block 8	
125	95
126	93
127	88
128	61
129	80
130	64
131	78
132	55
133	37
134	33
135	42
136	92
137	49
138	46
139	23
140	62
141	49
142	26
143	05
144	25
145	33

Item No.	Number of Students Out of 100
Block 11	
146	51
147	67
148	75
149	46
150	49
151	61
152	85
153	48
154	34
155	39
156	25
157	37
158	20
159	06

Item No.	Number of Students Out of 100
Block 13	
160	77
161	21
162	49
163	31
164	81
165	63
166	31
167	24
168	59, 21, 14, 6

Figure 4
Student Performance Data: 1992 Writing ALS

Prompt Score Point Distribution
(% of Students Papers Scored at Each Score Point)

Prompt	Scores						Off Task/Omit
	1	2	3	4	5	6	
01W9	9.1	41.1	28.7	6.1	0.8	0.0	14.4
012W7	6.6	32.0	35.2	18.0	1.6	0.5	6.3
01W3	3.1	8.2	42.4	30.7	6.5	1.3	8.8
01W8	6.4	18.7	40.6	24.8	3.5	0.6	5.5
O12W4	5.9	7.7	43.5	26.3	4.9	0.7	11.1

Mean Scores for Prompts

Group A	$\bar{X}$	Group B	$\bar{X}$
(P) 012W10	2.48	(P) 01W9	2.40
(N) 01W6	2.89	(N) 012W7	2.76
(I) 01W5	3.31	(I) 01W3	3.36
(P) 01W11	2.54	(N) 01W8	3.02
(I) 012W4	3.21	(I) 012W4	3.21

P=Persuasive
N=Narrative
I=Informational

Figure 5
Student Performance Data: 1998 Civics ALS

Item	Percent Correct	Mean	Percentage						
			1	2	3	4	Omit	Not Reached	Off-Task
1	83								
2	76								
3	43								
4	58								
5	65								
6		1.99	33	31	32		3	0	0
7	50								
8	36								
9	62								
10	32								
11	70								
12	43								
13		2.04	21	48	24		5	1	1
14		2.52	7	28	53		10	1	1
15	56								
16	23								
17	67								
18	53								
19	54								

Figure 6.1
Interrater Consistency Feedback for the Basic Level: 1992 Math ALS

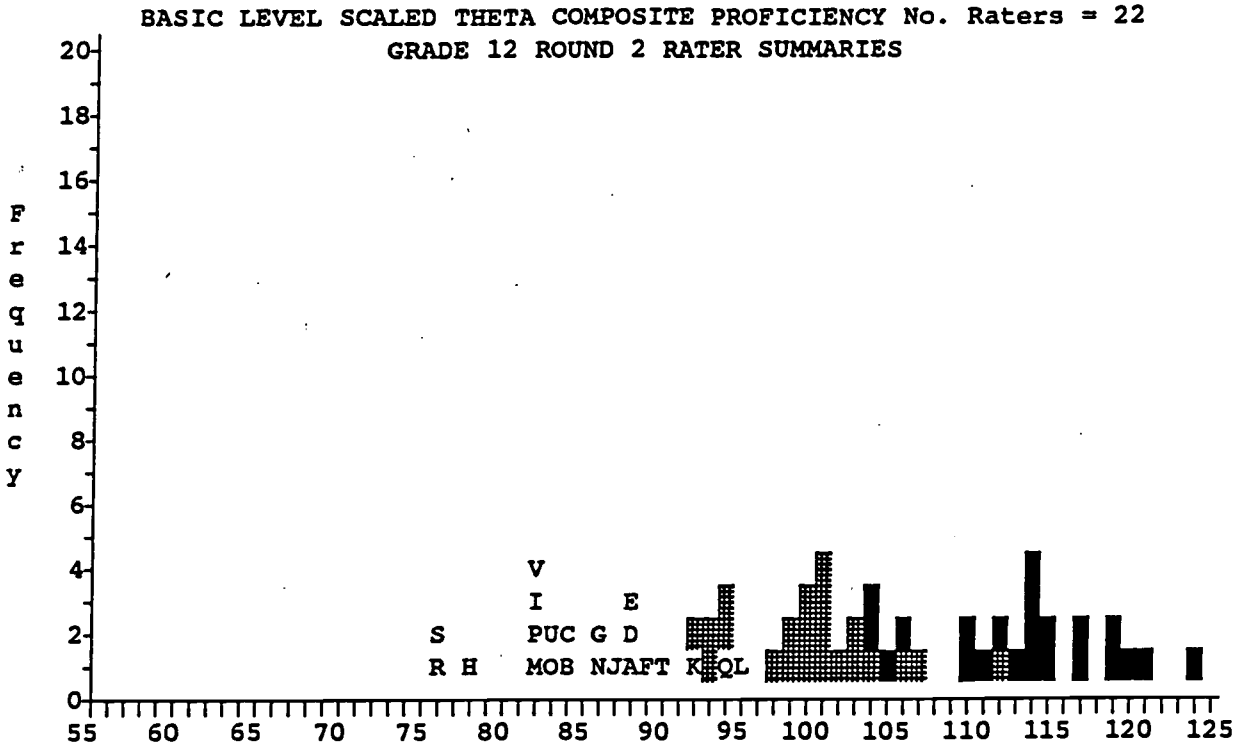


Figure 6.2
Interrater Consistency Feedback for the Proficient Level: 1992 Math ALS

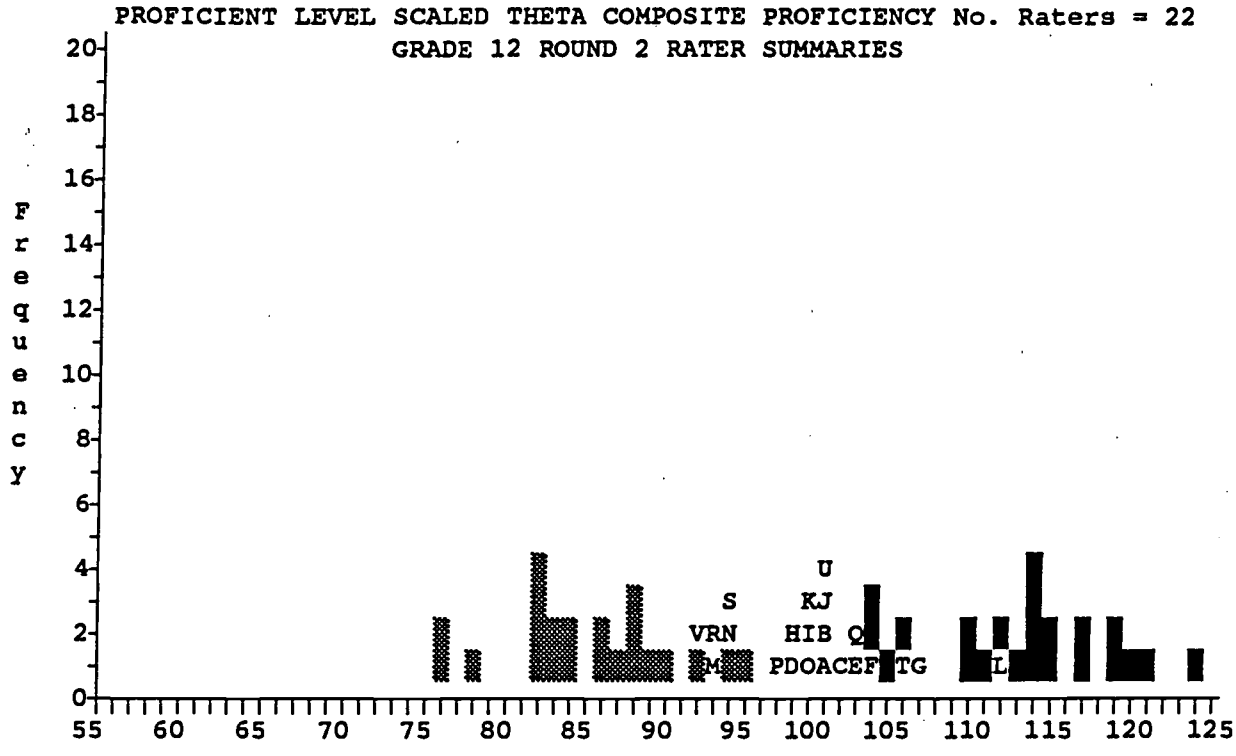


Figure 6.3
Interrater Consistency Feedback for the Advanced Level: 1992 Math ALS

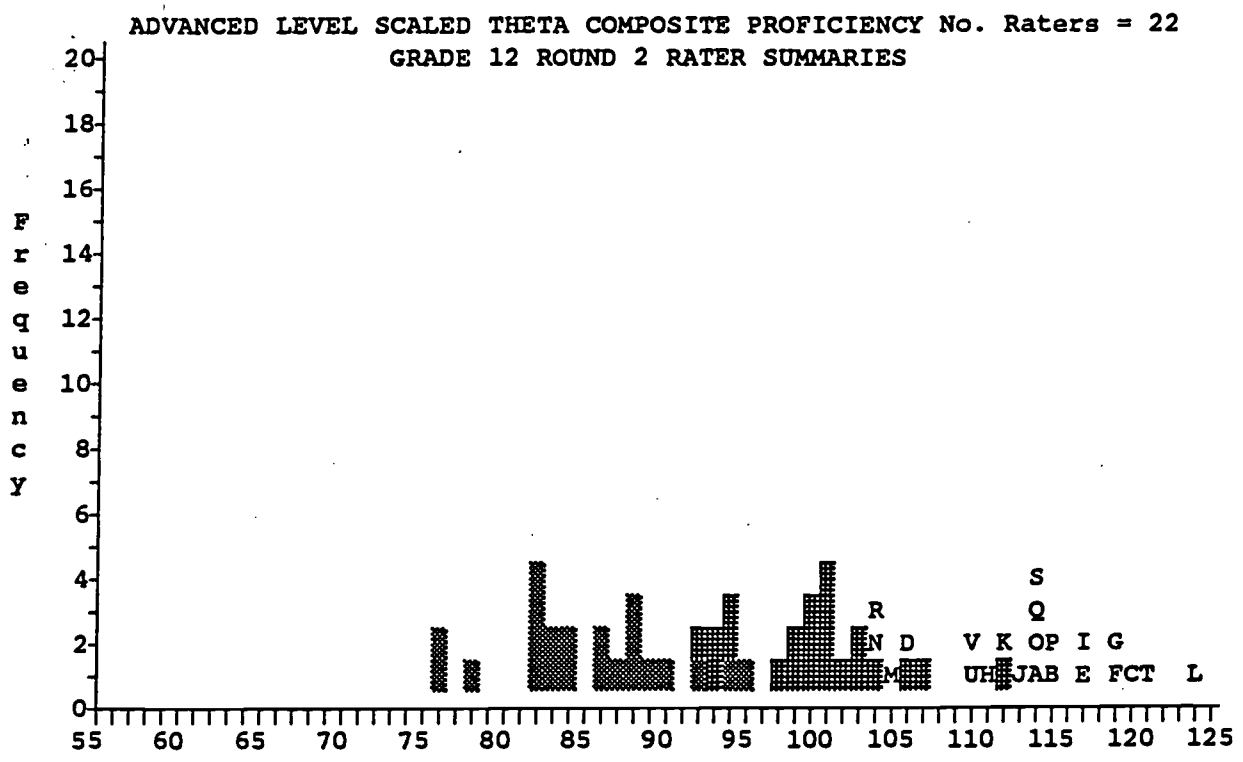
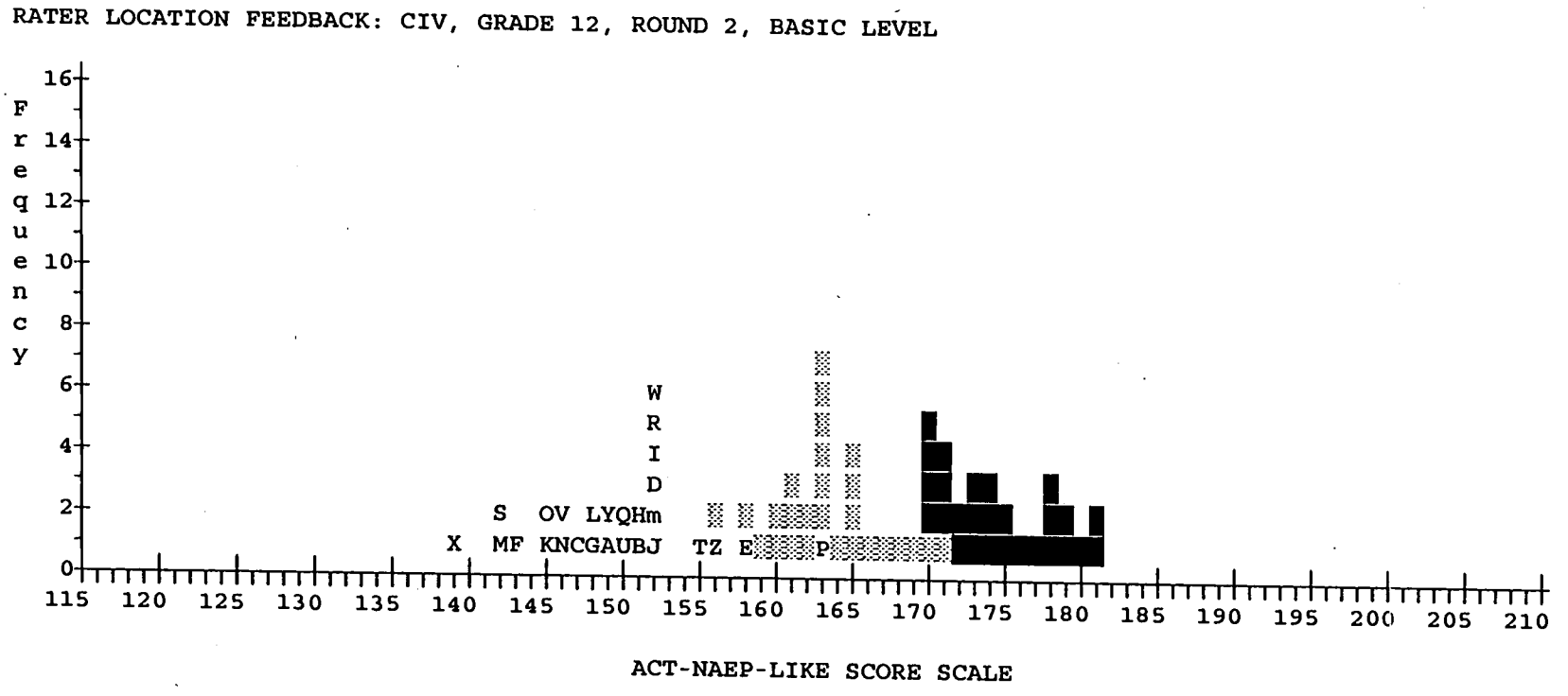
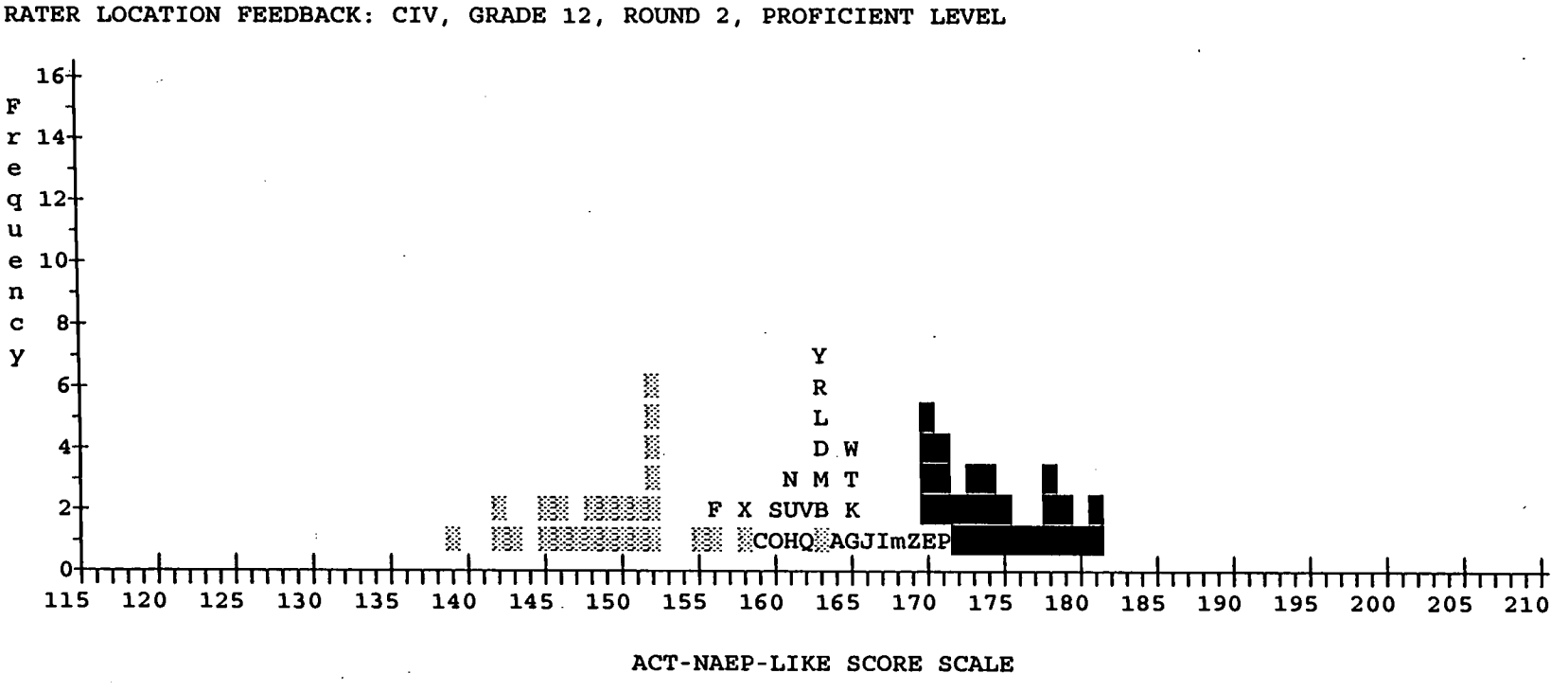


Figure 7.1
Interrater Consistency Feedback for the Basic Level: 1998 Civics ALS



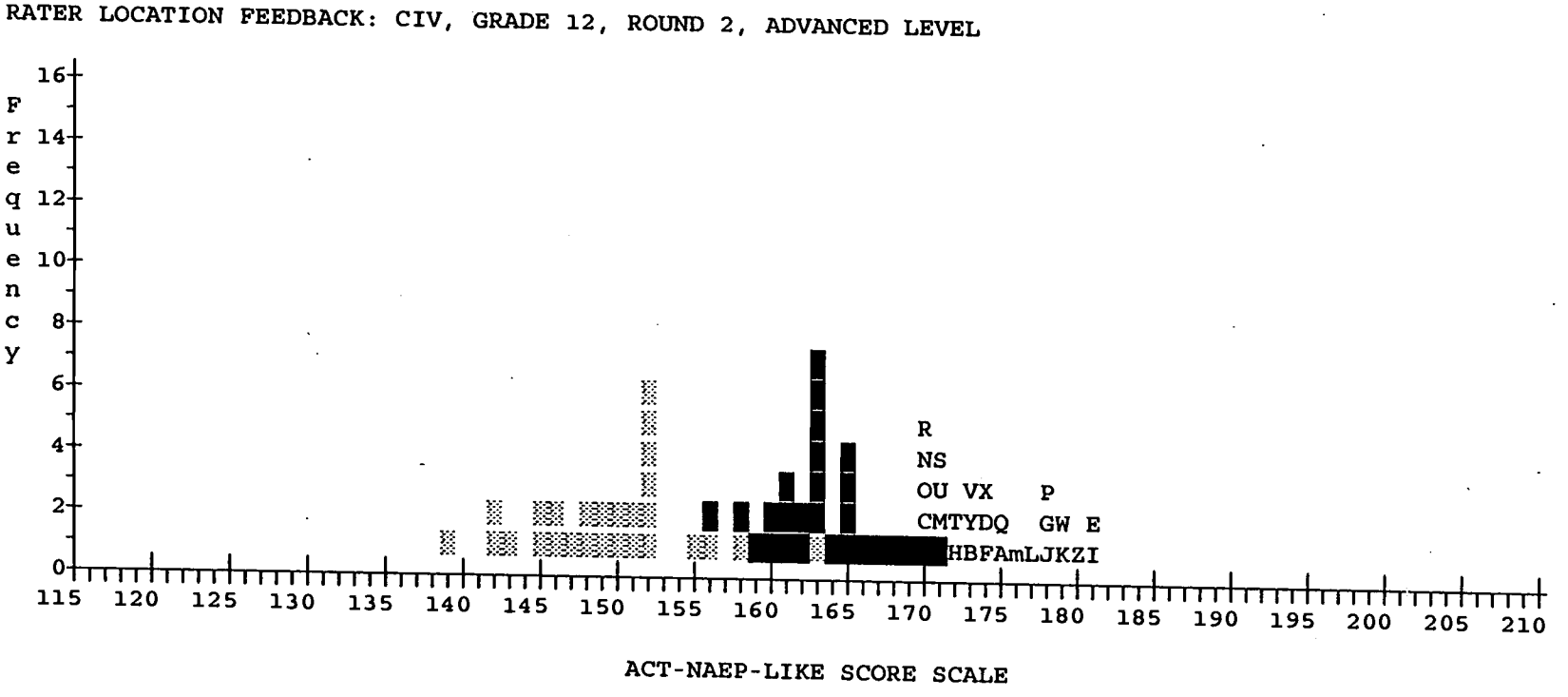
Legend: A-Z are secret IDs of panelists.
▒ represents rater location for Proficient Achievement Level.
■ represents rater location for Advanced Achievement Level.

Figure 7.2
Interrater Consistency Feedback for the Proficient Level: 1998 Civics ALS



Legend: A-Z are secret IDs of panelists.
▒ represents rater location for Basic Achievement Level.
▮ represents rater location for Advanced Achievement Level.

Figure 7.3
Interrater Consistency Feedback for the Advanced Level: 1998 Civics ALS



Legend: A-Z are secret IDs of panelists.
▒ represents rater location for Basic Achievement Level.
■ represents rater location for Proficient Achievement Level.

Figure 8
Whole Booklet Feedback: 1994 Geography ALS

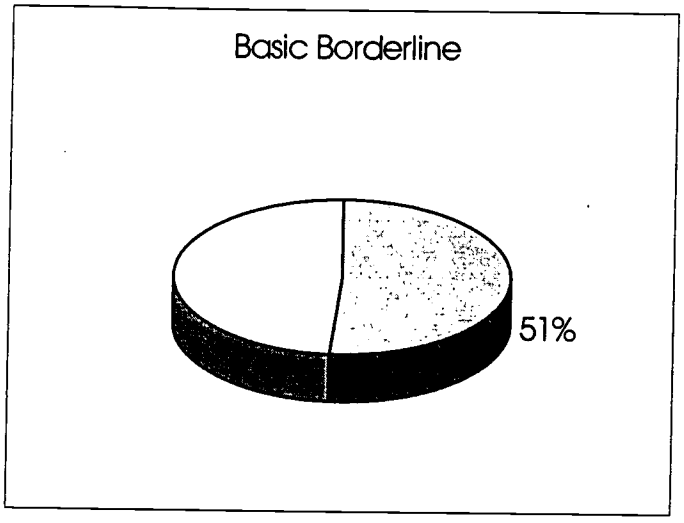
Geography ALS Study: Round 3
Whole Booklet Exercise Data Grade 04, Group A

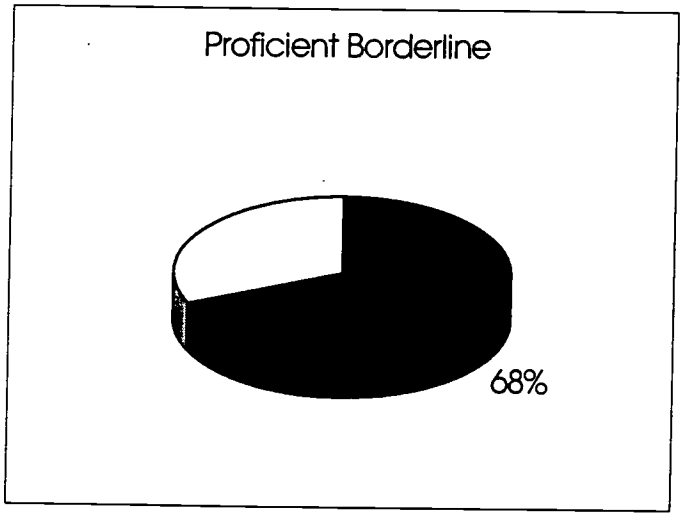
Based on your grade's average ratings, students performing at the borderline Basic Level are expected to get 52.3% correct on this booklet.

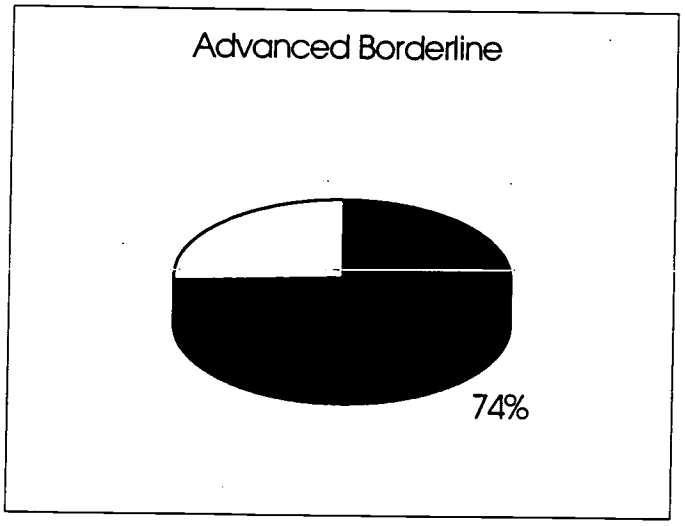
Based on your grade's average ratings, students performing at the borderline Proficient Level are expected to get 75.9% correct on this booklet.

Based on your grade's average ratings, students performing at the borderline Advanced Level are expected to get 88.7% correct on this booklet.

Figure 9
Whole Booklet Feedback: 1996 Science ALS

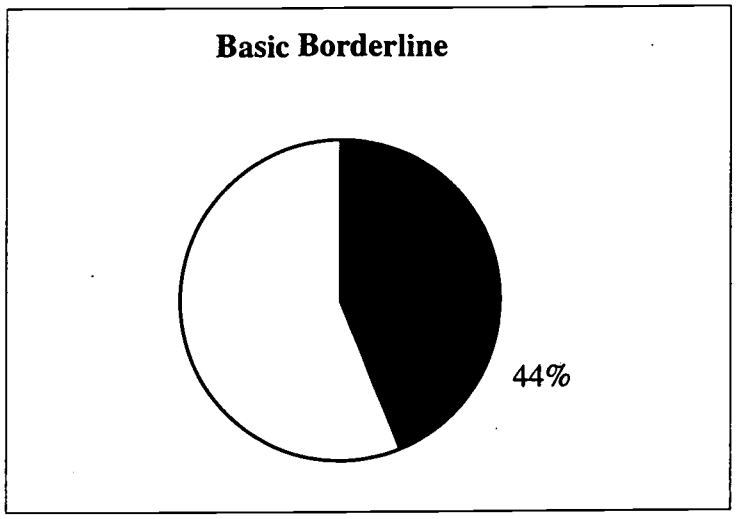


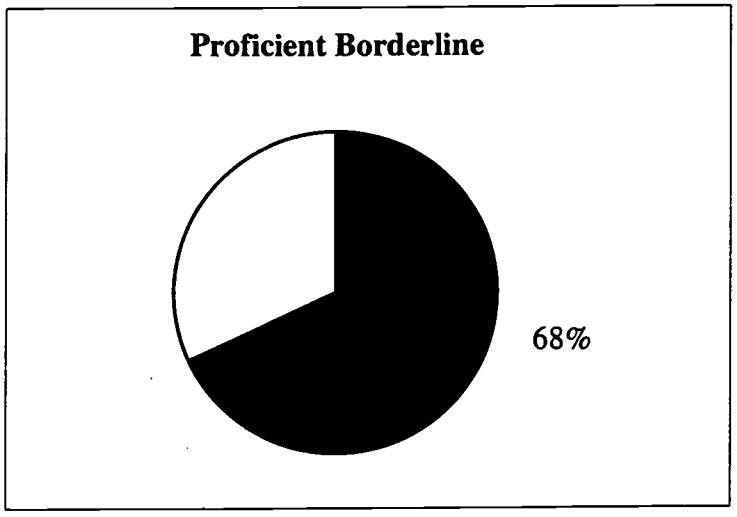


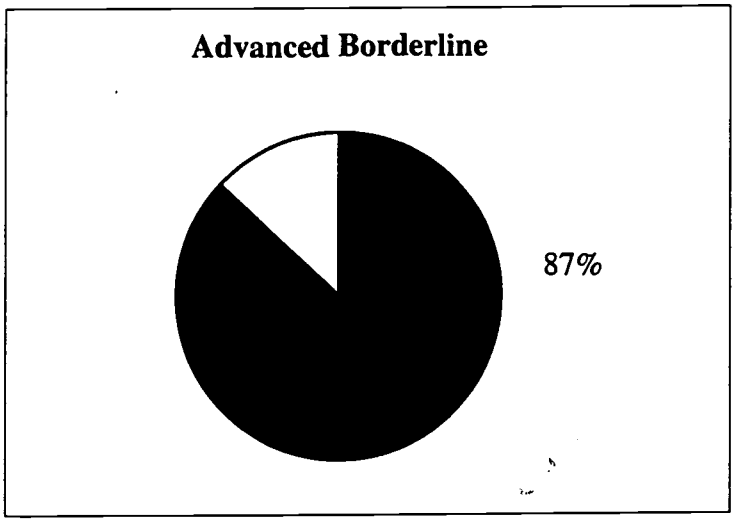


These pie charts report information to you for the test booklet form that was used when you took the NAEP exam. The percents reported on the charts are the percents of total possible points that students performing at the borderline of each achievement level are expected to get correct. These percents are estimated for the particular items in the NAEP test booklet, given where your grade set the achievement level cutscores in this round.

Figure 10
Whole Booklet Feedback: 1998 Writing ALS







These pie charts report information for the NAEP test booklet you completed on day one. The charts show the percent of total possible points that students performing at the borderline of each achievement level would need to earn. The percents were estimated using the achievement levels cutpoints that your group set in this round.

Figure 11
Intrarater Consistency Feedback: 1992 Writing ALS

Panelist ID	Prompt	Basic Level		Proficient Level		Advanced Level	
		Mean	Range	Mean	Range	Mean	Range
X04X	W10	3.50	1	1.50	1	2.50	3
	W6	1.00	0	2.00	0	5.50	1
	W5	1.50	1	4.00	2	2.50	1
	W11	1.00	0	4.00	0	5.00	0
	W4	1.50	1	3.50	3	4.00	0
Rater Average=		1.70		3.00		3.90	

Figure 12
Intrarater Consistency Feedback: 1994 Geography ALS

Geography ALS Study: Grade 04, Round 3
Intrarater consistency feedback information for rater GX040X

Your overall rating average for Basic Level performance is 146.1.
Given that, your ratings for these items are relatively low,

Q1G3 7	KJ000706	101.8
Q12G7 9	BO001946	101.8
Q12G8 3	BO001968	101.8
Q12G811	KJ000588	101.8
Q1G6 2	SE000717	123.4

and, your ratings on these items are relatively high.

Q12G8 7	KJ000863	168.3
Q1G6 7	KJ000761	165.8
Q12G713	BO001951	160.8
Q12G813	BO001954	158.3
Q12G7 4	BO001939	158.3

Your overall rating average for Proficient Level performance is 163.0.
Given that, your ratings for these items are relatively low,

Q1G3 9	BO001933	150.8
Q1G6 2	SE000717	150.8
Q1G313	KJ000704	151.7
Q1G312	KJ000747	153.3
Q12G7 2	BO001936	154.2

and, your ratings on these items are relatively high.

Q12G8 7	KJ000863	179.9
Q1G3 7	KJ000706	175.0
Q12G714	BO0001952	174.1
Q1G6 7	KJ000761	174.1
Q12G811	KJ000588	173.3

Your overall rating average for Advanced Level performance is 177.7.
Given that, your ratings for these items are relatively low,

Q1G313	KJ000704	168.3
Q1G614	BO001930	168.3
Q1G615	KJ000698	168.3
Q12G8 2	BO001967	169.1
Q12G712	BO001949	169.5

and, your ratings on these items are relatively high.

Q1G3 7	KJ000706	208.2
Q12G7 9	BO001946	204.9
Q12G8 6	SE000692	201.6
Q12G8 7	KJ000863	191.6
Q1G6 7	KJ000761	191.6

Figure 13
Intrarater Consistency Feedback in the 1998 ALS Process: Reckase Charts

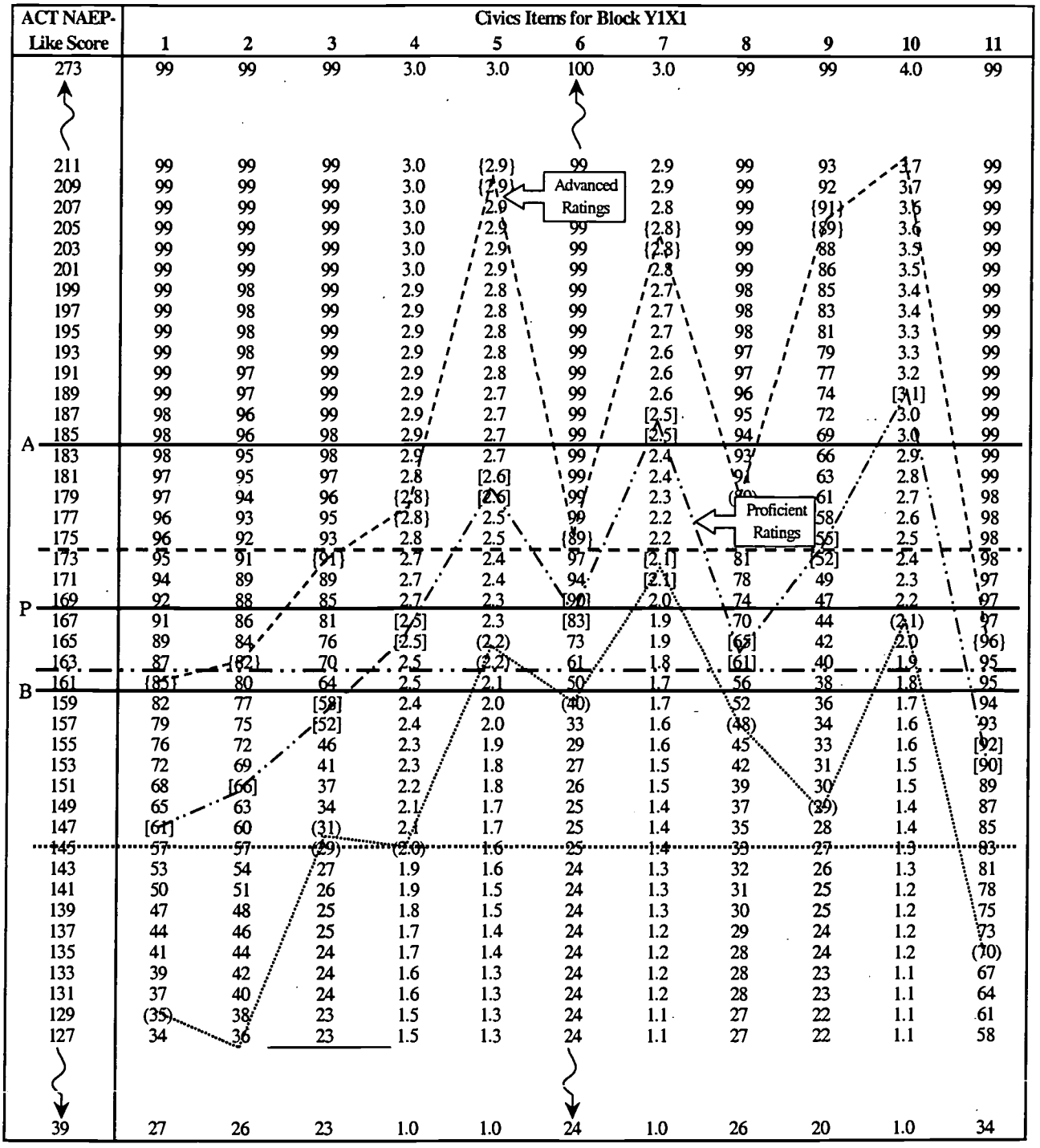


Figure 14
Consequences Data Feedback in the 1994 ALS Process

Geography ALS Study: Round 3
Cutpoints and SDs for Grade 04

	Basic	Proficient	Advanced
Cutpoints:	148.3	166.5	178.6
SDs:	7.5	4.3	4.0

Performance Distribution Data for Grade 04

		Rating Based
Percentages at or above Basic Level	=	73.3%
Percentages at or above Proficient Level	=	22.9%
Percentages at or above Advanced Level	=	4.9%

Figure 15
Consequences Feedback Data in the 1996 Science ALS Process

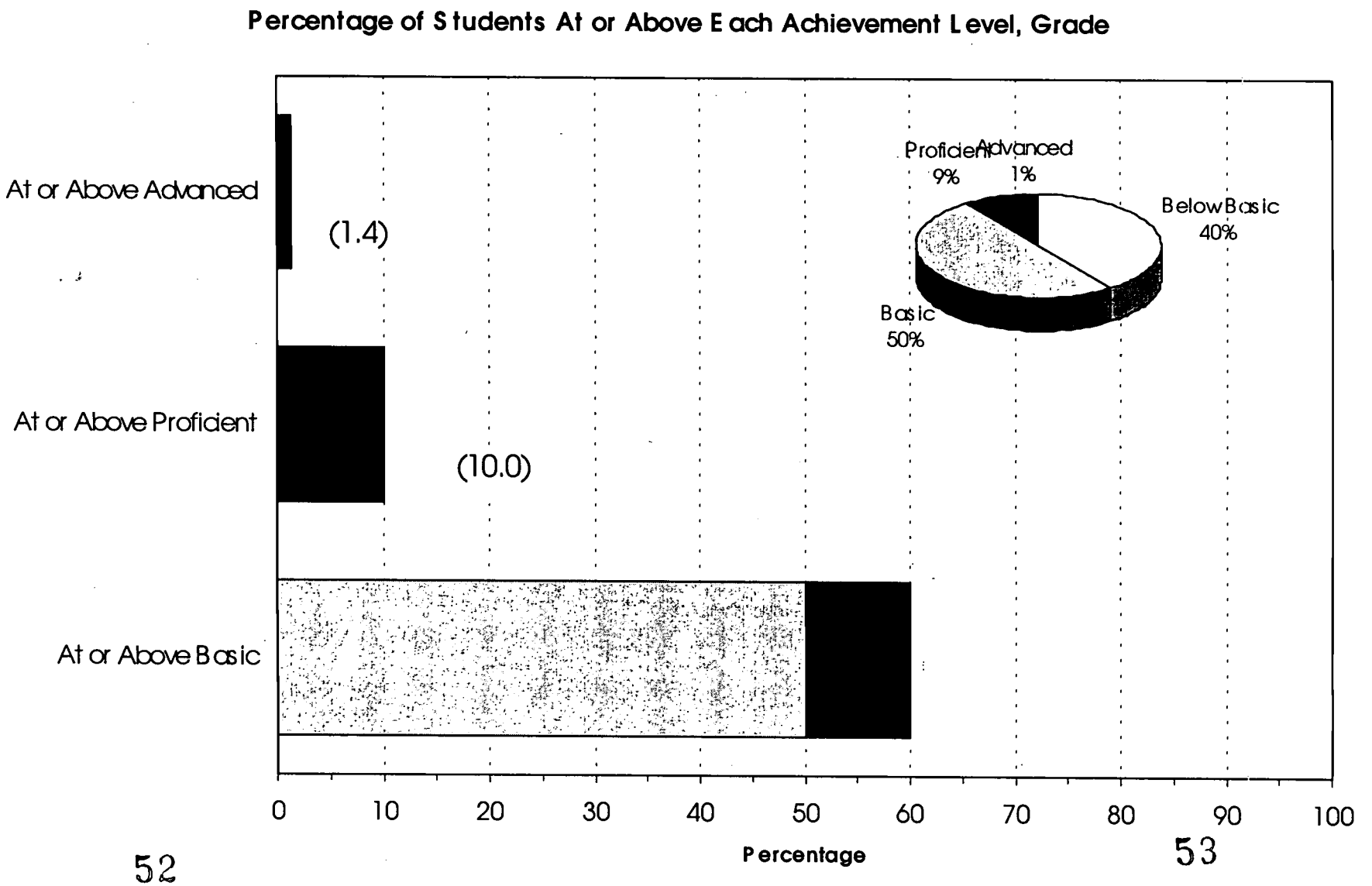
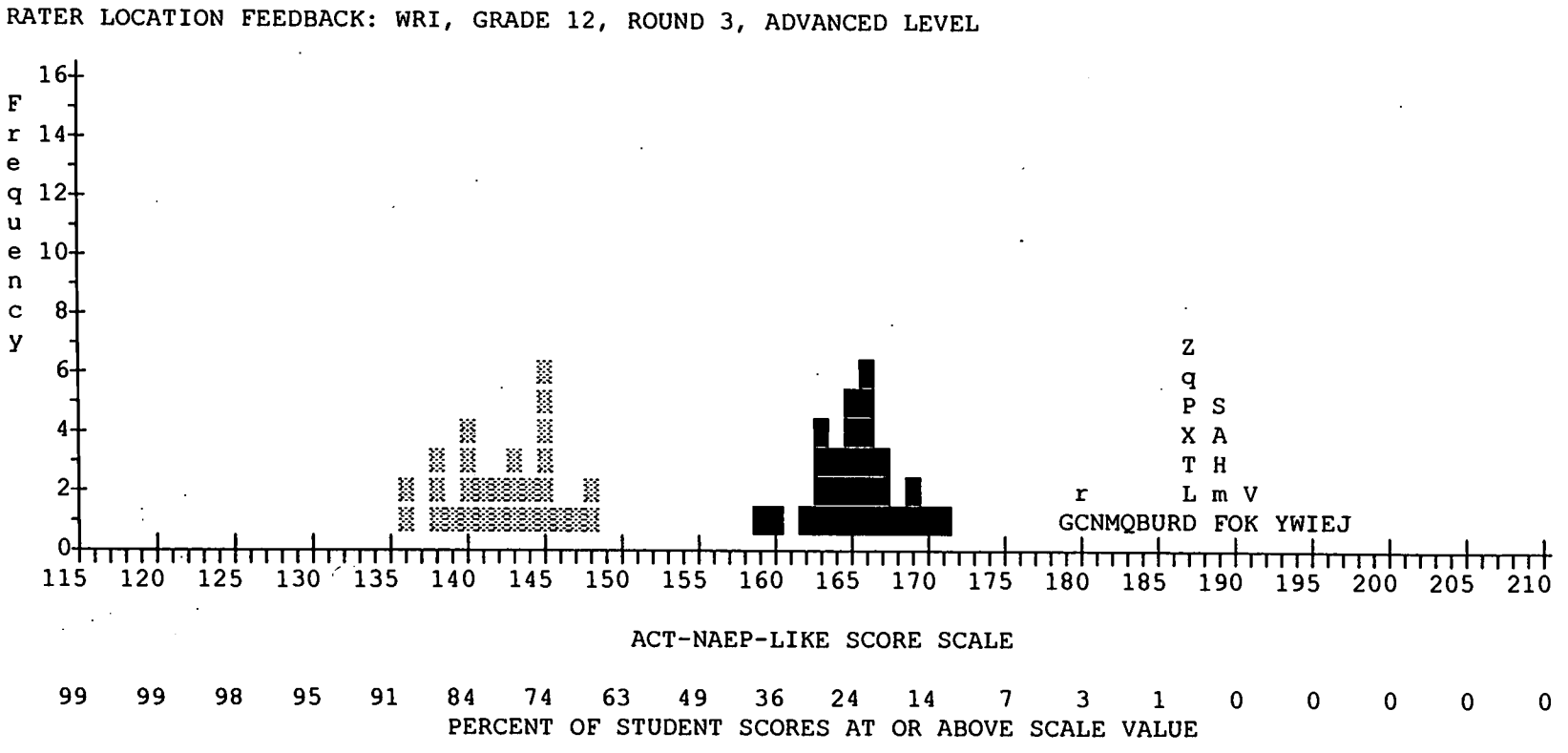


Figure 16
Consequences Data Reported on Round 3 Rater Location Charts: 1998 Writing ALS



Legend: A-Z are secret IDs of panelists.
▒ represents rater location for Basic Achievement Level.
■ represents rater location for Proficient Achievement Level.

Figure 17
Individual Consequences Feedback for Round 3: 1998 Writing ALS

GROUP CONSEQUENCES: GRADE 12

	BASIC	PROFICIENT	ADVANCED
%>=	80.1	22.5	0.8

INDIVIDUAL RATER CUTPOINTS AND CONSEQUENCES DATA: GRADE 12

RATER ID	BASIC		PROFICIENT		ADVANCED	
	Cutpoint	(%>=)	Cutpoint	(%>=)	Cutpoint	(%>=)
A:	148.2	(67.6%)	169.5	(15.5%)	189.9	(0.4%)
B:	143.7	(77.7%)	160.9	(34.0%)	184.2	(1.6%)
C:	145.8	(73.2%)	164.9	(24.3%)	180.1	(3.6%)
D:	143.3	(78.4%)	164.1	(26.2%)	187.9	(0.7%)
E:	147.2	(70.4%)	170.4	(13.4%)	196.1	(0.1%)
F:	145.2	(74.1%)	166.7	(20.8%)	189.4	(0.6%)
G:	145.9	(73.2%)	166.5	(20.8%)	179.9	(3.6%)
H:	145.9	(73.2%)	166.0	(21.7%)	189.4	(0.6%)
I:	141.1	(82.9%)	167.2	(19.2%)	195.0	(0.1%)
J:	146.3	(72.3%)	171.3	(12.2%)	197.1	(0.0%)
K:	142.2	(80.7%)	166.1	(21.7%)	191.0	(0.3%)
L:	141.7	(81.5%)	165.2	(23.4%)	187.4	(0.8%)
M:	140.1	(84.3%)	159.6	(37.3%)	182.5	(2.4%)
m:	148.3	(67.6%)	166.3	(20.8%)	189.9	(0.4%)
N:	138.3	(87.5%)	162.6	(30.1%)	181.8	(2.6%)
O:	138.8	(86.3%)	163.4	(28.2%)	190.7	(0.3%)
P:	144.6	(75.9%)	166.2	(21.7%)	187.9	(0.7%)
Q:	140.6	(83.7%)	163.4	(28.2%)	183.1	(2.0%)
q:	139.4	(85.7%)	163.4	(28.2%)	187.2	(0.9%)
R:	144.5	(75.9%)	167.1	(19.2%)	186.9	(1.0%)
r:	136.8	(89.2%)	165.3	(23.4%)	180.1	(3.6%)
S:	142.5	(80.1%)	165.9	(22.5%)	189.6	(0.4%)
T:	145.5	(74.1%)	164.4	(25.3%)	187.6	(0.8%)
U:	140.9	(82.9%)	166.8	(20.0%)	185.8	(1.1%)
V:	140.4	(83.7%)	167.6	(18.4%)	191.3	(0.3%)
W:	145.7	(73.2%)	169.9	(14.8%)	194.0	(0.1%)
X:	143.5	(78.4%)	165.5	(22.5%)	187.6	(0.8%)
Y:	136.5	(89.7%)	168.9	(16.3%)	193.4	(0.2%)
Z:	138.3	(87.5%)	163.2	(28.2%)	187.2	(0.9%)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

NOTICE

REPRODUCTION BASIS

☐ This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☑ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").